Hindawi
Security and Communication Networks
Volume 2022, Article ID 4976848, 14 pages
https://doi.org/10.1155/2022/4976848

Research Article

Multisemantic Path Neural Network for Deepfake Detection

Nan Wu, Xin Jin, Qian Jiang, Puming Wang, Ya Zhang, Shaowen Yao, and Wei Zhou

Engineering Research Center of Cyberspace, Yunnan University, Kunming 650091, Yunnan, China

Correspondence should be addressed to Qian Jiang; jiangqian_1221@163.com

Received 14 May 2022; Revised 6 August 2022; Accepted 27 September 2022; Published 11 October 2022

Academic Editor: Beijing Chen

With the continuous development of deep learning techniques, it is now easy for anyone to swap faces in videos. Researchers find that the abuse of these techniques threatens cyberspace security; thus, face forgery detection is a popular research topic. However, current detection methods do not fully use the semantic features of deepfake videos. Most previous work has only divided the semantic features, the importance of which may be unequal, by experimental experience. To solve this problem, we propose a new framework, which is the multisemantic pathway network (MSPNN) for fake face detection. This method comprehensively captures forged information from the dimensions of microscopic, mesoscopic, and macroscopic features. These three kinds of semantic information are given learnable weights. The artifacts of deepfake images are more difficult to observe in a compressed video. Therefore, preprocessing is proposed to detect low-quality deepfake videos, including multiscale detail enhancement and channel information screening based on the compression principle. Center loss and cross-entropy loss are combined to further reduce intraclass spacing. Experimental results show that MSPNN is superior to contrast methods, especially low-quality deepfake video detection.

1. Introduction

Automated video editing techniques have made great strides in the past few years with the development of deep learning. In particular, people have shown great interest in face manipulation. It is now easy to transfer facial expressions from one video to another based on generative adversarial networks (GANs) and autoencoders [1]. Even those who do not know deep learning can easily change one person's face to another in a few minutes [2], and a fake face is difficult for human eyes to distinguish. It is easy to change who the speaker is or what is said. While deepfake techniques bring benefits, there are hidden dangers.

These techniques open a new window for film and television. For example, dead movie stars can reappear through face manipulation, and people who do not exist in the real world can be created through GANs. Moreover, malicious attacks and revenge porn are a small part of malicious face manipulation. This also influences politics, such as by tampering with speech content and spreading fake news [3]. As a result, deepfake videos have attracted the interest of researchers, and methods to detect whether a face has been manipulated have become paramount.

Deepfake videos can have at least three levels of forgery characteristics: microscopic, mesoscopic, and macroscopic. Microscopic features correspond to unseen differences, such as anomalies in small regions. Macroscopic or semantic features refer to the whole image semantics that the human eyes can feel. Mesoscopic features are seen in between. Afchar et al. [4] designed MesoNet to detect mesoscopic features. Current deepfake video detection methods do not take full advantage of these three levels of features. Usually, authenticity discrimination has been based only on high semantic features, and the performance needs improvement. It is possible to design a network that can integrate the three levels for deepfake detection. However, semantic segmentation methods that can ensure the improvement of accuracy have not yet been proposed. Similar work was based on the practical experience of feature hierarchy division. In addition, it is uncertain whether the weights of the three hierarchical features are the same.

Deepfake video detection methods have achieved accuracy of nearly 100% for high-quality videos, but their accuracy for low-quality videos with high-compression rates needs to be improved [5]. For example, the accuracy of Xception [6] was 99.26% for the uncompressed FaceForensics++ dataset [7], but 72.93% at the C40 compression rate without pretraining. The high-compression rate makes the video very blurry and the forgery trace becomes unclear and not obvious, thereby becoming more difficult to distinguish the real video from the fake. Most videos on the Internet are compressed due to upload size limitations; thus, low-quality video forgery detection is significant. For this kind of video, we studied the commonly used H.264 video compression format, which includes inter and intraframe compression [8]. If only the original adjacent frames are removed through interframe compression, the accuracy will indeed be improved in theory. However, this will lead to inconsistencies with the creation requirements of benchmark datasets such as FaceForensics++, so we only use intraframe compression, which preserves the Y channel information on the YCrCb space and compresses the CbCr information as much as possible. Figure 1 shows the changes of different channels in an image at different compression rates. After comparative experimental analysis, we find that when the Y channel of the image is used as the input, the accuracy is higher than that when other channels are used. In addition, to highlight the high-frequency information of low-quality videos, multiscale detail enhancement was performed on images before channel separation. Based on the above two findings, we propose a deepfake detection method integrating different semantics in the network. We find no standard for semantic division from the aspect of channel level, but division from the aspect of the receptive field of the convolution kernel is reasonable.

When considering semantic level importance, instead of assigning weights manually, we use channel-spatial attention to assign them automatically. Therefore, a multichannel network with different receptive fields is proposed to integrate the features at different levels to capture forgery features. In constructing the neural network, the essential information is extracted through preprocessing and input to the network. We connect the feature maps of multiple pathways or semantics and automatically assign the weights to the three semantics through the channel-spatial attention module, perform feature fusion, and classify. We train and test our model on FaceForensics++ and DeepFake-TIMIT [9] and perform cross-validation on Celeb-DFv2 [10]. Experimental results show that our network has better accuracy than current methods, especially in low-quality deepfake video encoding.

This work makes the following contributions:

(1) A multiscale detail enhancement method is introduced in deepfake detection. Fuzzy features are extracted from three Gaussian kernels, the residuals are calculated with the original image, and the detailed texture features of the forged image are highlighted;

(2) Based on the study of video compression methods, the extraction of significant channel information assists in the detection of forged images with high-compression rates;

(3) A multipath network for multisemantic information fusion is proposed. The three kinds of semantic information are automatically assigned weights by a channel-spatial attention module, and low, medium, and high semantic information of forged images can be effectively divided and interpreted;

(4) Our method is evaluated on manipulated videos datasets. It performs well on the DeepFake-TIMIT and FaceForensics++ datasets and generalizes satisfactorily on Celeb-DFv2. The proposed preprocessing method can improve the detection of low-quality counterfeit videos, and the network can comprehensively capture different semantic information of images.

2. Related Work

We summarize current fake video generation methods, analyze deepfake detection methods, and introduce our method.

2.1. Deepfake Image Generation. Image generation techniques have developed rapidly over the past two decades, and methods such as StyleGAN [11] can produce fake images or videos that are credible to the eye. It is especially difficult to see traces of forgery after a video is compressed. Juefei-Xu et al. [12] produced a comprehensive report on counterfeiting generation and detection. Deep learning generation techniques of deepfake videos include autoencoders and GANs. Forgery methods can be categorized by the generated results as entire face synthesis, attribute manipulation, identity swap, and expression swap, as shown in Figure 2. Entire face synthesis generates a face that does not exist in the world. The input of these networks is a random vector, and the output is a realistic fake face image. Many models can be used, such as WGAN [13], StyleGAN, and PGGAN [14]. Attribute manipulation can modify the attributes of a person's head, including simple attributes such as expression, hair color, and baldness and complex attributes such as gender, age, and the wearing of glasses. Classic examples are StarGAN [15] and STGAN [16]. Identity swapping, which replaces a face in a source image with a target's face, has attracted much interest in recent years. Apps such as Zao [17] allow one to swap identities with a favorite star. Moreover, there are malicious attacks. Examples of identity swapping methods include FaceSwap [18] and CycleGAN [19]. Also known as face reconstruction, face-swapping is somewhat similar to identity swapping, replacing the source image's facial expression with that of the target image's facial expression, which include Face2Face [1] and A2V [20].

Methods of forgery generation include AAMS [21] for style transfer, SC-FEGAN [22] for image repair, and SAN [23] for super-resolution, but most of these methods are not the focus of face manipulation detection. According to the risk rank, identity swapping entails the most risk, followed by expression swap. Entire face synthesis and attribute manipulation are not very dangerous.

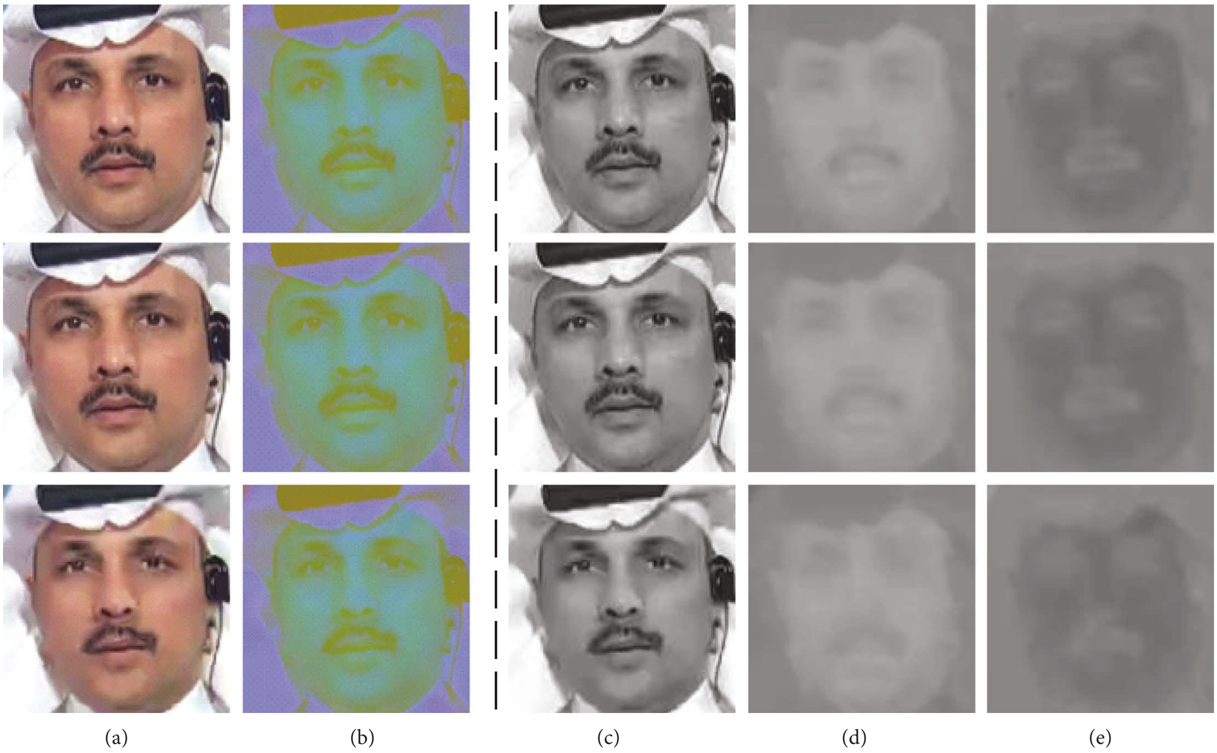

FIGURE 1: Changes in YCrCb channels of images with three compression ratios (declining video quality): (a) the image under the RGB channel; (b) the result of conversion to YCrCb channels; (c), (d), and (e) Y, Cr, and Cb images, respectively. Changes of Cr and Cb channel information are most apparent.

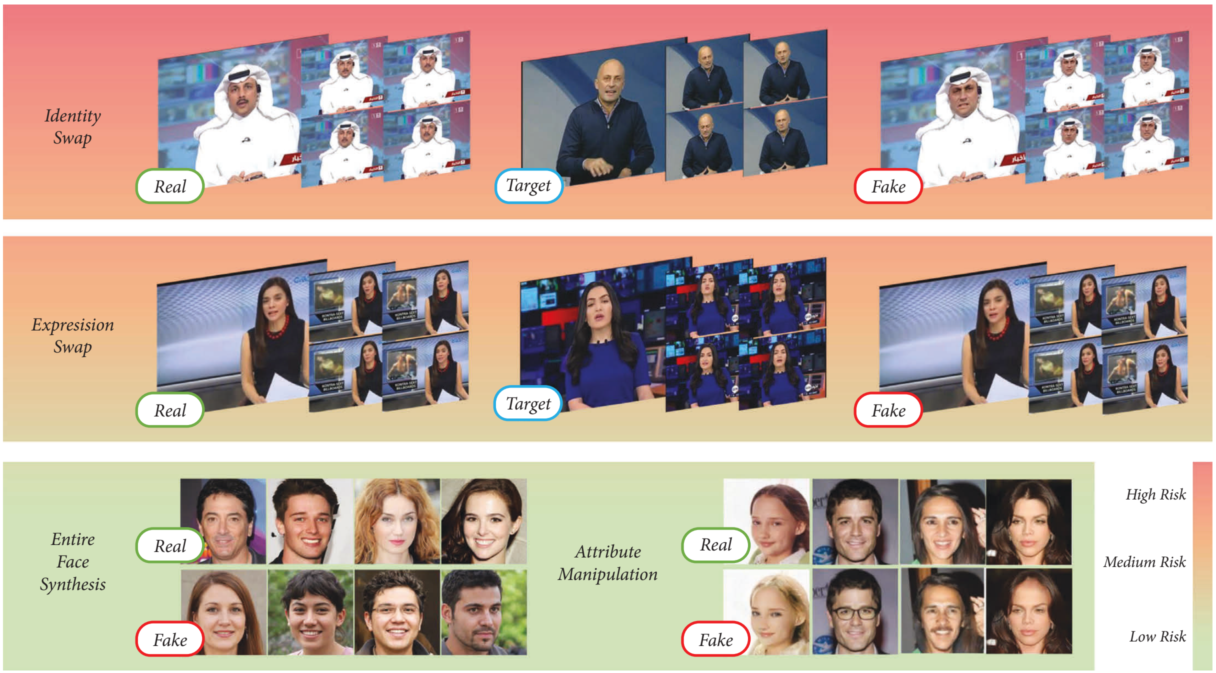

FIGURE 2: Samples of four forgery categories. Colors at lower-right indicate risk level.

2.2. Deepfake Image Detection. Methods to detect deepfake features are based on spatial or image pixels, the frequency domain, or biological signals. Spatial-based methods use either conventional feature forensics or deep learning. Conventional image forensics relies on specific manipulation evidence [24], using frequency domain and statistical features such as local noise analysis, illumination, and device fingerprints to distinguish deepfakes. Nataraj et al. [25]

extracted co-occurrence matrices on three color channels in the pixel field and conducted classification training according to these features. Although the conventional forensics technique is mature, several shortcomings are present in dealing with deepfake videos because it pays more attention to abnormal features of local images. Deepfake videos are usually processed to avoid detection, such as by compression methods, compression rates, and condensation. Therefore, the conventional feature forensics technique cannot be directly applied to detect deepfake videos.

Methods based on or combined with deep learning have recently gained attention [26–29]. Sabir et al. [30] used recurrent neural networks to capture temporal differences in fake videos. Liu et al. [31] conducted an empirical study on real and fake faces and obtained some important findings. One of these findings is that the texture of a fake face is fundamentally different from that of a real face. Deep learning techniques and large datasets make it easier to catch the features associated with forgery [32]. This method can judge the authenticity of a single-frame image and detect video frames by a combined strategy, but it has limitations. Most learning models rely on the same dataset with the same data distribution for both training and testing and are weak in the face of unknown tampering types [33]. At the same time, the ability of deep learning models to detect highly compressed video frames is greatly reduced.

The method based on the frequency domain analyzes the differences of deepfake images such as through a Fourier or wavelet transform [34]. Durall et al. [35] proved that standard upsampling methods lead the forged images generated by these models to fail and to correctly reproduce the spectral distribution of natural training data. Most methods calculate feature maps with the differences between true and fake images in the frequency domain, and combine deep learning such as the support vector machine (SVM) for classification. Because the available spectrum of high-resolution images is much smaller than that of high-resolution photos, it is challenging to identify compressed videos.

Biometric authentication techniques have developed in recent years [36]. Detection methods based on biological signals cannot reproduce natural physiological characteristics by using fake videos, and the physiological characteristics of fake faces are inconsistent with those of real faces. [37]. Therefore, biological signal detection-based methods are constantly being developed by researchers. For example, by monitoring minimal periodic changes in skin color, Qi et al. [38] speculated that the normal heartbeat rhythm would be interrupted by deepfakes and proposed a dual temporal attention network. Although detection methods based on physiological signal characteristics can effectively make use of the defects of deepfake techniques, these methods gradually become invalid with the continuous improvement of generation methods, such as the addition of physiological characteristics (e.g., blink frequency). Besides, methods based on hard-to-find biological signals, such as heart rate, would be far less accurate due to video compression and other processing [39].

Because conventional forensic techniques are easily avoided by new deepfake techniques, frequency domain feature-based statistical methods are not strong at detecting low-resolution forged videos, and biological signal-based methods are weak in improving generation technique. Most current work still adopts data-driven deep learning methods. As far as we know, current deep learning methods do not fully use the three semantics of images. For example, Mesonet only used mesoscopic semantics, while later networks used macroscopic semantics for judgment, such as Xception [7], FDFtnet [40], and AMTEN [41]. Zhao et al. [42] used microscopic and macroscopic semantics. Although some previous work mentioned semantics, they could not explain the relationship between network depth and the three types of semantics. Our work developed a targeted solution to this problem; specifically, the three semantics are set according to the width of the network, which has better interpretability. Moreover, ablation experiments show that the proposed method is effective and can surpass current methods at detecting forged images, especially in low-resolution videos. In addition, according to the compression principle, we propose a preprocessing method for low-resolution video.

3. Proposed Method

Based on the above analysis, we design a multisemantic path neural network (MSPNN) for deepfake detection to capture deepfake features under different semantics, as shown in Figure 3.

3.1. Multiscale Detail Enhancement. We use a multiscale approach to enhance the details of the source image. We first define three Gaussian filters:

$$G_1 = \begin{bmatrix} 0.0030 & 0.0133 & 0.0219 & 0.0133 & 0.0030 \\ 0.0133 & 0.0596 & 0.0983 & 0.0596 & 0.0133 \\ 0.0219 & 0.0983 & 0.1621 & 0.0983 & 0.0219 \\ 0.0133 & 0.0596 & 0.0983 & 0.0596 & 0.0133 \\ 0.0030 & 0.0133 & 0.0219 & 0.0133 & 0.0030 \end{bmatrix}, \tag{1}$$

$$G_2 = g_2 * g_2^{\mathrm{T}},$$

where $g_2 = [0.02760.06630.12380.18020.20420.18020.12380.06630.0276]$ and

$$G_3 = g_3 * g_3^{\mathrm{T}}, \tag{2}$$

where $g_3 = [0.00810.01370.02200.03300.04650.06160.07660.09000.10150.09000.07660.06160.04650.04650.03300.02200.01370.0081]$.

Then, we obtain three fuzzy images using Gaussian image filters

$$\begin{aligned} B_1 &= G_1 \otimes I_{in}, \\ B_2 &= G_2 \otimes I_{in}, \\ B_3 &= G_3 \otimes I_{in}, \end{aligned} \tag{3}$$

where G_1, G_2, and G_3 are Gaussian kernels with respective kernel sizes of 5 × 5, 9 × 9, and 19 × 19 and standard

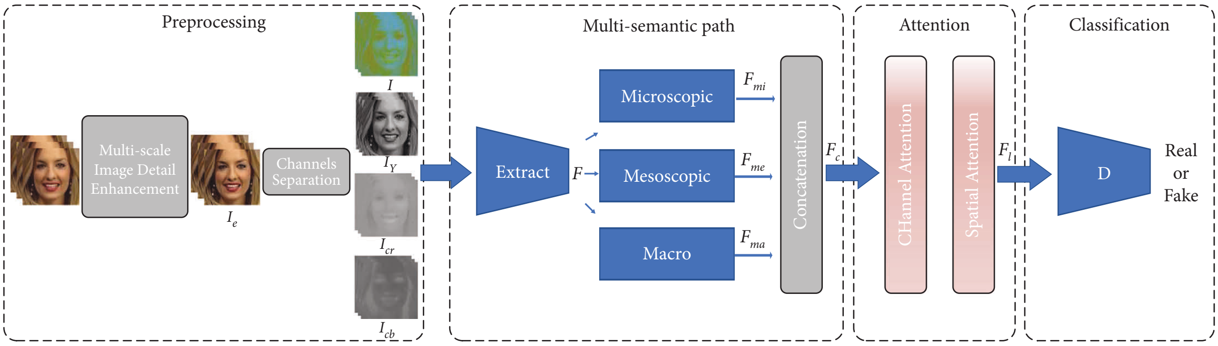

FIGURE 3: Cropped images of faces are used as input. After multiscale image detail enhancement preprocessing, I_e is obtained, and YCrCb channel separation is performed. The Y channel image I is taken as input. Preliminary feature F is extracted and put into the microscopic, mesoscopic, and macroscopic semantic channels. *Fmi*, *Fme*, and *Fma* are obtained by feature extraction of the three channels. These are fused into the channel and spatial attention modules, and the weight of the three semantic feature maps is allocated. Results are input to the discriminator for classification.

deviations $\sigma_1 = 1.0$, $\sigma_2 = 2.0$, and $\sigma_3 = 4.0$; $\otimes$ represents convolution; and B_1, B_2, and B_3 are the three filtered images. The fine, intermediate, and coarse details are, respectively, extracted as

$$\begin{aligned} D_1 &= I_{in} - B_1, \\ D_2 &= B_1 - B_2, \\ D_3 &= B_2 - B_3. \end{aligned} \tag{4}$$

We combine the three layers to generate a detailed image of the whole:

$$D^* = (1 - w_1 \times \operatorname{sgn}(D_1)) \times D_1 + w_2 \times D_2 + w_3 \times D_3 + I_{in}. \tag{5}$$

According to experience, ω_1, ω_2, and ω_3 are fixed as 0.5, 0.5, and 0.25, respectively. Figure 4 shows the process of image detail enhancement. Figure 5 shows the effect of multiscale detail enhancement. Faces at the top in Figure 5 are slightly blurred, while at the bottom, detail enhancement makes the visual perception of local details clearer, which aids in the detection of forged images with high compression.

3.2. Compressed Videos Analysis.

3.2. Compressed Videos Analysis. According to our research, the detection accuracy of high- and medium-quality deepfake videos, i.e., uncompressed and medium-compressed, respectively, is close to 100%, while that of high-compression videos is much worse, especially for some videos with more realistic tampering effects. Therefore, research on high-compression forged video must be improved. Since human eyes are not sensitive to the chromaticity of an image but are sensitive to its brightness, during image compression, it is desirable to retain as much chromaticity information as possible and compress brightness information to save storage space. Since the chrominance information of the compressed video hardly changes, the definition of the video does not change significantly. Since compression is carried out in YCrCb color space and our datasets are RGB images, spatial conversion is first required, given as follows:

$$\begin{bmatrix} Y \\ Cr \\ Cb \end{bmatrix} = \begin{bmatrix} 0.299 & 0.578 & 0.114 \\ 0.500 & -0.4187 & -0.0813 \\ -0.1687 & -0.3313 & 0.500 \end{bmatrix} \times \begin{bmatrix} R \\ G \\ B \end{bmatrix} + \begin{bmatrix} 0 \\ 128 \\ 128 \end{bmatrix}, \tag{6}$$

where R, G, and B are the gray values of the three components of RGB.

Figure 1 shows images with different compression rates. The compression rate increases gradually from the first to the third row. The first line is the original image, and the image that is almost visually lossless in the second row is slightly compressed. The third row is a low-quality image. Column (a) shows images in RGB color space, and column (b) shows images under the YCrCb channels. Column c, column d, and column e show separate images using the YCrCb channels, such as the Y channel, the Cr channel, and the Cb channel, respectively. The change in the Y channel is the least obvious, and the change in the Cr and Cb channels is the most obvious. Inspired by the above observations, we extract the image information of the RGB channel into two types of luminance information and one type of chrominance information, i.e., the YCrCb channel. Then, we conducted four experiments using the Y channel, the Cr channel, the Cr channel, and the original image separately to verify our idea. Experimental results show that using only Y channel information can improve the accuracy of highly compressed video and has little effect on slightly compressed video.

3.3. Multisemantic Path. MSPNN can output feature maps with multiple semantics through different receptive fields and network depths. The features of these different layers are finally connected, and a learnable weight is added to the three feature layers for fusion classification. The final classification relies on the deep feature map and considers the shallow and middle feature maps. The overall framework is shown in Figure 3.

The network has three parts. First is simple image preprocessing to generate 32 feature maps. Different feature maps are generated through three semantic channels. The network details are shown in Figure 6. Since low semantics can be understood as microscopic images, all filters in the semantic channel adopt a 3 × 3 window. The high semantics are the macroscopic features of the image, and the corresponding receptive field is more extensive, so the filter size of the semantic channel is 7 × 7. Inspired by Inception [43], we replace a 7 × 7 convolutional kernel with three 3 × 3 convolutional kernels, which can reduce computation without reducing the receptive field and can have more nonlinear transformations, as shown in Figure 6. Mesoscopic semantics is between mesoscopic and macro semantics. The receptive field of this channel is 5 × 5, and we use two 3 × 3 convolution kernels. Considering the influence of network depth semantics, the three semantic depths are also increased.

3.4. Semantic Integration. Although the microscopic, mesoscopic, and macroscopic semantics of images are juxtaposed, their importance is not the same. Hence, we apply a weight to each of the semantics instead of feeding back directly to the discriminator. In our model, these weights are learnable, which we accomplish through a channel-attention module to combine space and channels; this can achieve better results than SENet [44], which only pays attention to the channel. The first one is the channel-attention module of the image given as follows:

$$\begin{aligned} M_c(F) &= \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \\ &= \sigma\left(W_1\left(W_0\left(F^c_{avg}\right)\right) + W_1\left(W_0\left(F^c_{\max}\right)\right)\right), \end{aligned} \tag{7}$$

where σ denotes the sigmoid function, $W_0 \in \mathbb{R}^{C/r\times C}$ and $W_1 \in \mathbb{R}^{C/r\times C}$. Note that MLP weights W_0 and W_1 are shared for both inputs, and ReLU activation is followed by W_0. Then the spatial attention is

$$\begin{aligned} M_s(F) &= \sigma\left(f^{7\times 7}([AvgPool(F); MaxPool(F)])\right) \\ &= \sigma\left(f^{7\times 7}\left(\left[F^s_{avg}; F^s_{\max}\right]\right)\right), \end{aligned} \tag{8}$$

where $f^{7\times 7}$ represents convolution with a 7 × 7 filter, and AvgPool() and MaxPool() are average and maximum pooling, respectively. The fused feature map is fed to the final classifier.

3.5. Loss Function. According to our investigation, the center loss function, while used in many face recognition tasks [45], does not improve performance in tasks such as handwritten number recognition. We conclude that the center loss function is more suitable for fine-grained classification tasks. To this end, we introduce a center loss function to our model as

$$Lc = \frac{1}{2}\sum_{i=1}^{m}\left\|x_i - c_{yi}\right\|_2^2, \tag{9}$$

where $c_{y_i} \in \mathbb{R}^d$ represents the distribution center of yi category data; that is, the feature center of true or fake faces, x_i represents the feature before the full connection layer, and m is the batch size. We use this loss to continually decrease the sum of squares of the distance between the feature maps of each sample and the feature, i.e., to make the in-class distance as small as possible.

Normally, c_{y_i} should be updated as the depth features change. The choice of feature centers should consider the entire training set and average the features of each class in each iteration. Specifically, c_{y_i} is updated in small batches, and the centers are calculated by averaging the characteristics of the corresponding classes in each iteration. Second, to avoid large disturbances caused by a small number of mislabeled samples, we use the scalar α, which is limited to the range [0, 1], to control the learning rate of the center. The updated equation of c_{y_i} is

$$\Delta \mathbf{c}_j = \frac{\sum_{i=1}^{m}\delta(y_i = j)\cdot\left(\mathbf{c}_j - \mathbf{x}_i\right)}{1 + \sum_{i=1}^{m}\delta(y_i = j)}, \tag{10}$$

where if $y_i = j$ is satisfied, then $\delta(y_i = j) = 1$; otherwise, $\delta(y_i = j) = 0$; that is, when the tags y_i and C_j are of different categories j, then C_j does not require updating. We use a cross-entropy loss function and central loss joint supervision to train the network to learn true and fake features. The equation of the final loss function is given as follows:

$$\mathscr{L} = \mathscr{L}_S + \lambda\mathscr{L}_C. \tag{11}$$

We first consider Ls and Lc of equation (11) equally important, so we set λ as 1. Values can have different effects on the result, and we believe that multiple attempts can find a more suitable value. We compute

$$Ls = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i\log\left(S\left(\hat{y}_i\right)\right) + \log\left(1 - y_i\right)\log\left(1 - S\left(\hat{y}_i\right)\right)\right], \tag{12}$$

where $\hat{y}_i$ is the score of the i-th face, and $y_i \in 0, 1$ is the related face label, where the label 0 is associated with faces from real, original videos, and 1 is associated with fake videos. N is the total number of faces used to train each batch, and $S(\cdot)$ is the sigmoid function.

4. Experimental Results and Analysis

We describe popular datasets, video segmentation methods, and their implementation, describe pretreatment ablation experiments and comparative experiments with other methods, and discuss verification of generalization.

4.1. Datasets. Our experiments use the FaceForensics++, DeepFake-TIMIT, and Celeb-DFv2 datasets. FaceForensics++ is one of the largest and most diverse deepfake datasets. It is a prominent face forgery dataset widely used in deepfake detection, with 1,000 YouTube videos. The authors of FaceForensics++ used four types of face tampering to create fake videos, including FaceSwap, DeepFakes, Face2-Face, and NeuralTextures. A total of 1000 deepfake videos are generated with each tampering method, including videos

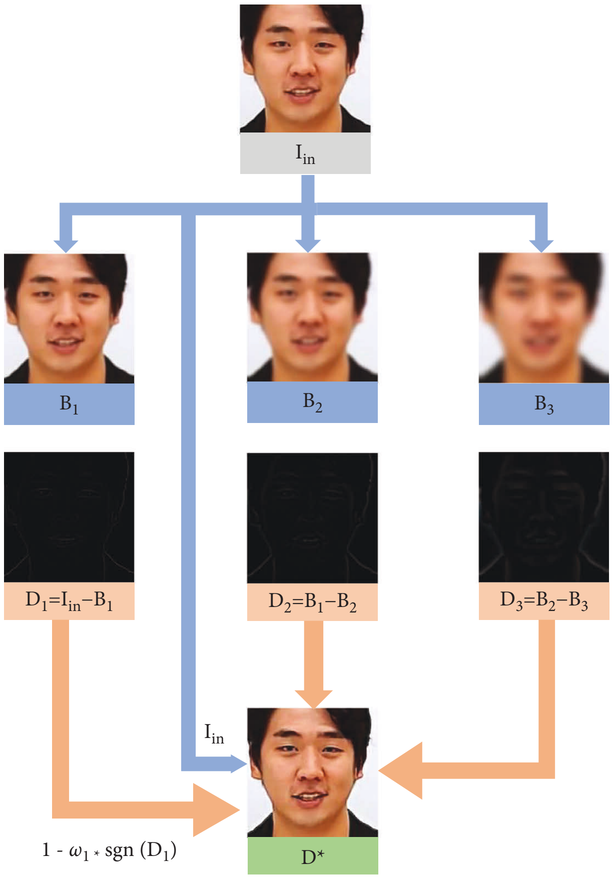

Figure 4: B_1, B_2, and B_3 are results of the three Gaussian filters. D_1, D_2, and D_3 are details of calculation of the original image and filtered results. The final image D^* is enhanced by incorporating details in the original image.

compressed with the original compression rate (C0), videos compressed with the micro compression rate (C23), and low-quality videos (C40). FaceForensics++ datasets have 1000 fake videos and 1000 real videos for each compression rate. When detecting forged videos, we divided the datasets into training, validation, and test sets according to the standard of FaceForensics++. There are 720 training sets, 140 validation sets, and 140 test sets.

DeepFake-TIMIT is generated by the face exchange algorithm based on the VidTIMIT dataset, which was developed using the faceswap-GAN method. Furthermore, Deepfake-TIMIT is the first deepfake dataset generated by GAN. The 640 generated fake videos are available in high (128×128) and low (64×64) quality. The production quality is better than that of Faceforensic++, but the video resolution is not high. We divided the dataset according to the settings of FaceForensics++. There are 320 videos of the two qualities, 230 training sets, 45 verification sets, and 45 test sets.

Celeb-DFv2 is a challenging deepfake video dataset that improves upon some weaknesses of other datasets. For example, UADFV, Faceforensic++, and Deepfake-TIMIT have low image resolution, poor quality of synthesized videos, rough tampering traces, and excessive flicker of video faces. The dataset consists of 590 real videos and 5,639 deepfake videos. Real videos from YouTube show celebrities of different genders, ages, and races.

For a fair comparison, we processed the video according to the clipping of FaceForensics++. All videos were framed, and dlib [46] was used to extract the feature points of each frame of the face to help locate and clip the face area, which was expanded by 1.3 times. Each video of the cropped face was taken in 30 frames. For data preparation of frame-level streams, we used OpenCV to extract frames. Since the datasets only operate on the faces in the video, not all frame information is helpful for deepfake detection from this perspective [7]. We focused our analysis on the area of the subject's face, and therefore on human faces, using dlib for face detection, which further reduced the amount of data processing. When extracting a face, dlib sometimes fails to recognize the face in a video frame, in whose case we skipped the frame and kept a constant number of faces captured in each video.

Figure7 shows the input image samples and output feature maps in the three experiments. The first line uses the low-compressed DeepFakes datasets in FF++ for training and testing. The generation method of forged image in the second line is the same as in the first line, with a higher compression rate. The third line uses the DeepFakes datasets with low compression in FF++ for training and Celeb-DFv2 for testing so as to verify the generalization performance. The output feature maps are the result of the fusion of the three paths. It can be seen from Figure 7 that the real image with higher brightness is concentrated in the center of the feature map, while the forged image with higher brightness is concentrated in the lower part.

4.2. Implementation. All experiments were performed on RTX 3090. The baseline [7] has a high accuracy in uncompressed datasets, and we only evaluated our model on low- and high-compressed data. We implemented MSPNN using the PyTorch deep learning library. For more details, we selected cross-entropy as the loss function in the training phase. The output of the network was distributed between 0 and 1, and we adopted the autoadaptive algorithm Adam in the optimization process. The initial learning rate was 1e-4, and the policies of cosine annealing LR were both used. The center loss function used the SGD optimizer. Batch normalization was used in each convolution to reduce the impact of overfitting. Dropout was introduced in the final full connection, with a ratio of 0.5. The batch size of the input data was 32. We trained our models with 100 epochs. The graph of the learning rate with each epoch was similar to a cosine function. The rest of the model settings were default values, the random seed was 43, and the input image was 224×224.

4.3. Preprocessing Analysis. Preprocessing had two steps. Multiscale detail enhancement highlights face textures, especially low-quality images, which are so blurred that it is difficult to see forged traces. In this process, three filters of different sizes, G_1, G_2, and G_3, were used to filter the image to obtain fuzzy images B_1, B_2, and B_3. The original image was

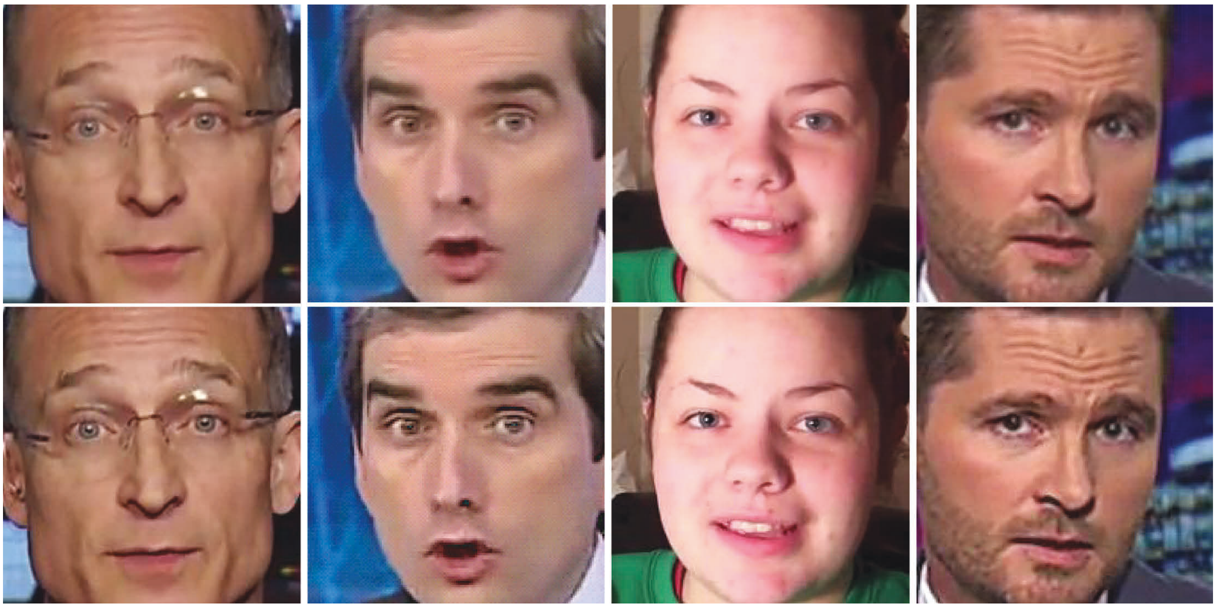

Figure 5: Effect of multiscale detail enhancement. The top part shows the original images and the bottom part shows images with clear texture after multiscale detail enhancement.

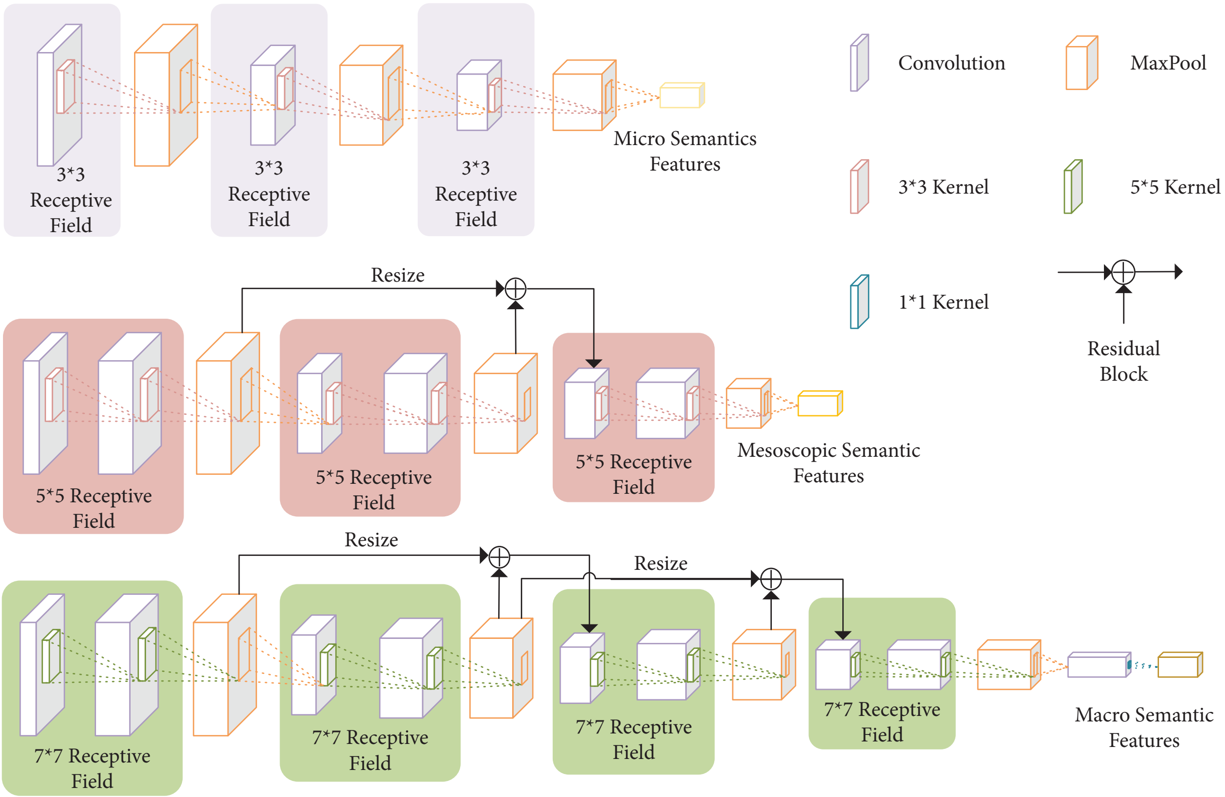

Figure 6: Network details of MSPNN. All receptive fields in micro semantic pathway are 3 × 3. Superposition of two 3 × 3 convolution kernels replaces 5 × 5 receptive fields in the mesoscopic semantic path, and skip connection is used in the second block. Two 5 × 5 convolution kernels replace 7 × 7 receptive fields in the macroscopic semantic path. In the third and fourth blocks, skip connections reduce loss of information. Finally, the output of each path is aligned with other semantic feature maps through downsampling.

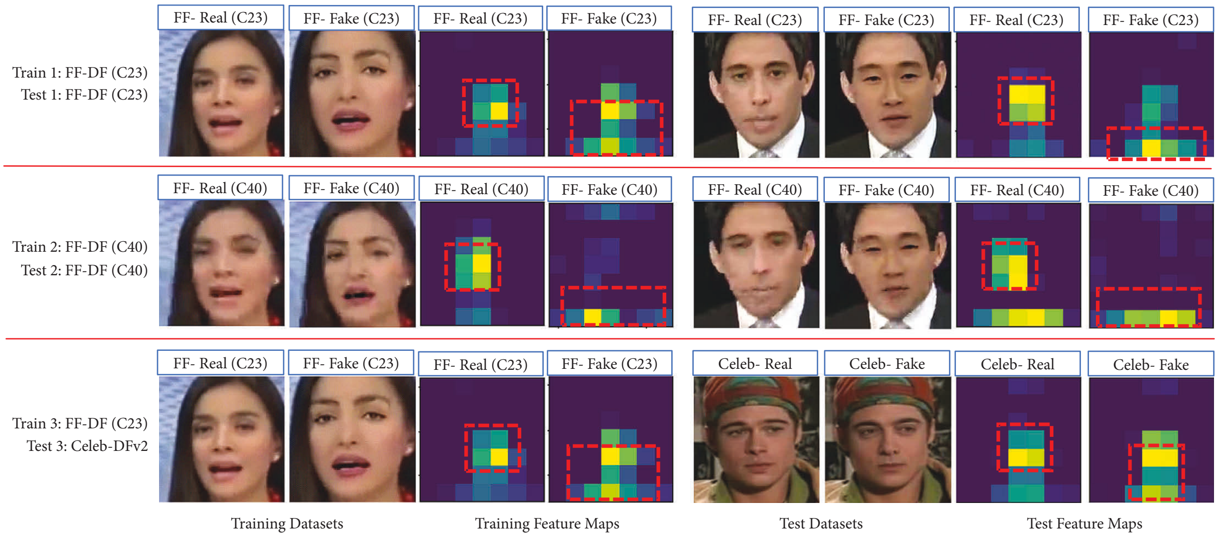

Figure 7: Training set, test set, and output feature maps. Red boxes indicate differences between real and fake images. It can be seen that the real image with higher brightness is concentrated in the center of the feature map, while the forged image with higher brightness is concentrated in the lower part of the feature map.

Table 1: Accuracy comparison of multiscale detail enhancement methods for datasets with different compression rates before and after introduction.

	Acc (%) on FF++(HQ)				Acc (%) on FF++(LQ)			
	DeepFakes	Face2Face	FaceSwap	NeuralTextures	DeepFakes	Face2Face	FaceSwap	NeuralTextures
Without detail enhancement	**99.73**	**99.09**	**99.17**	91.2	93.83	91.15	**92.47**	74.41
With detail enhancement	99.54	98.92	98.86	**91.3**	**94.25**	**91.25**	91.74	**74.52**

The bold values indicate the better results in the two experiments.

subtracted from B_1 to obtain detail image D_1. The detail image D_2 was obtained by combining detail image D_1 and fuzzy image B_2, and the detail image D_3 was obtained by combining detail image D_2 and fuzzy image B_3. The three detail images were fused with the original image to enhance the detail images. The improved results are shown in Figure 5. Ablation experiments were performed on the datasets of FaceForensics++ with compression rates C23 and C40, as shown in Table 1, from which we can see that the detection performance of the high-compression dataset was effectively improved compared with the low-compression dataset, which shows the effectiveness of the proposed preprocessing method for low-quality datasets. It is worth noting that the proposed detection was improved at any compression rate on the most challenging NeuralTextures dataset. The proposed method only modifies the facial expression corresponding to the mouth, leaving the eye area unchanged, and requiring more subtle detection methods.

The second preprocessing step was channel separation for high-compression images with low detection accuracy. We investigated the video compression standard H.264 and found that the measure keeps the information of the Y channel as much as possible while compressing the other two channels. In Figure 1, we can see the changes in the knowledge of the three channels after compression. So we converted the RGB image to a YCrCb image, and the images of Y, Cr, and Cb channels were taken out for training. We found that the accuracy of the image containing the brightness information channel is much higher than that containing the chroma information channel. The accuracy of the chromaticity information channel is much lower than of that of the ordinary RGB channel, as shown in Table 2, according to which most subset accuracy can be improved by using only Y channel information on the FaceForensics++ dataset, especially on the highly compressed C40 dataset. The experimental effect on some datasets becomes worse, but this change is not very large. We believe that the forged image with a low compression rate is close to the original image, so the effect is not apparent.

4.4. Experimental Results. Most detection methods are based on macroscopic semantics, i.e., the final feature maps of the network. The difference between a natural face and a fake is often subtle and occurs in the local area. Minor artifacts caused by the deepfake method are usually stored in the shallow characteristic of texture information. We believe that the microscopic semantic or superficial semantic features cannot be ignored. Focusing only on details is also flawed. A microscopic analysis based on image noise cannot be applied to the compressed video environment, where the image noise is strongly reduced. It is difficult for the human eye to distinguish the forged images at the same higher semantic level, especially in fine-grained analyzes, such as face discrimination. Therefore, our work takes into account the three kinds of semantic information, which receptive

Table 2: Y, Cr, Cb, and RGB channels used as input for training and test results for C23 and C40 datasets.

	Acc (%) on FF++(HQ)				Acc (%) on FF++(LQ)			
	DeepFakes	Face2Face	FaceSwap	NeuralTextures	DeepFakes	Face2Face	FaceSwap	NeuralTextures
RGB	**99.73**	99.09	99.17	**91.2**	93.83	**91.15**	92.47	74.41
Cr channel	84.46	85.58	79.04	83.35	79.67	75.73	75.32	63.32
Cb channel	87.19	85.17	77.08	80.11	80.45	72.91	69.65	59.83
Y channel	99.57	**99.21**	**99.34**	91.08	**94.35**	91.01	**92.55**	**74.92**

Bold values indicate the best results for four different channels.

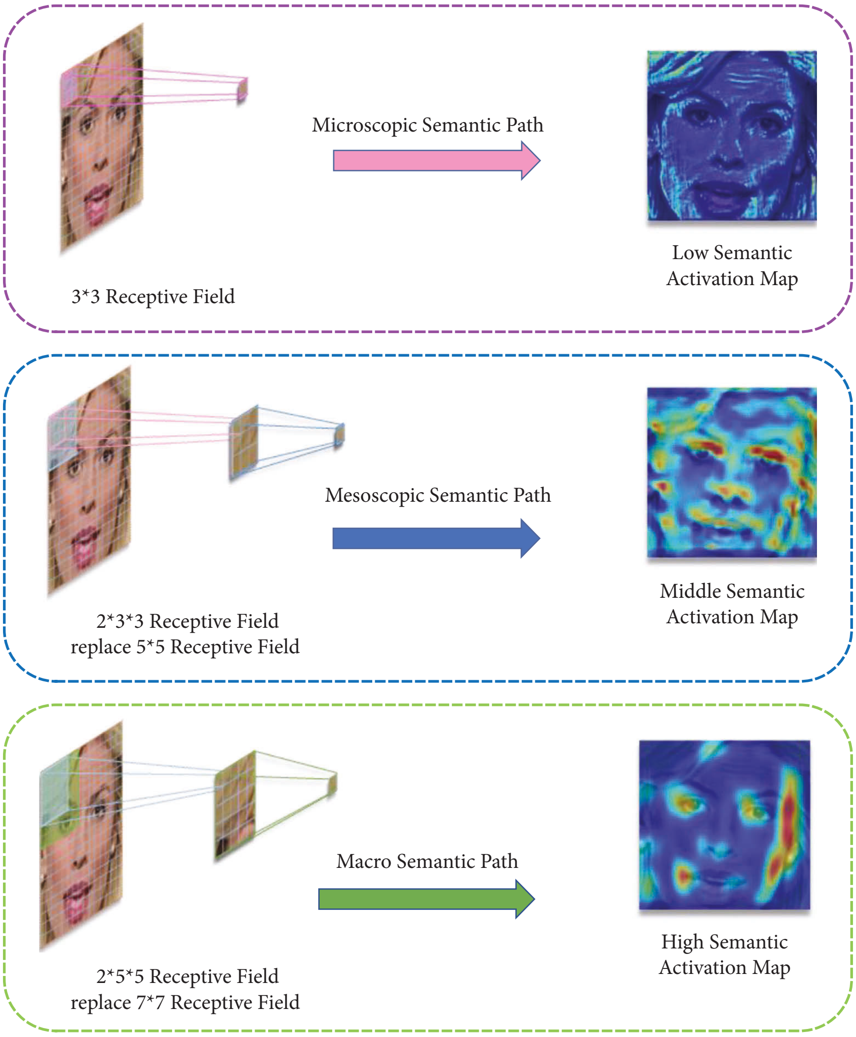

Figure 8: Receptive field description maps of different semantic paths and generated results.

fields of various sizes can also capture, and which are more explanatory, as shown in Figure 8.

Gradient-weighted class activation mapping is used to visually display the details of the attention of the three pathways, and it is evident that the microscopic semantic pathway pays more attention to details and the mesoscopic semantic path to multiple blocks. Macro semantics focus more on areas that are difficult to forge, such as the eyes, nose, and mouth because these are the most difficult to reproduce during the generation of forged images. In addition, small convolution kernels are used to map large convolution kernels, which reduces the computation of convolution, while increasing multiple nonlinear activations, and the receptive field is unchanged. We add residual blocks to the mesoscopic and macroscopic semantic pathways to ensure that information is not lost when the network depth increases. In Table 3, we can observe that much of the accuracy of the datasets is improved under the three pathways. They have a poor effect on some datasets, in particular the NeuralTextures dataset, which only tampers with parts of the images, whereas our microscopic semantic pathway captures much information that is not helpful to the detection of these datasets. Our addition of preprocessing makes up for this problem, as shown in Table 1. We also

Table 3: Comprehensive precision comparison of three semantic pathways on FF++ dataset, and comprehensive precision comparison of three semantic pathways and multiple semantic pathway networks on neuraltextures, deepfakes, FACE2FACE, faceswap, and neuraltextures.

	Acc (%) on FF++ (HQ)				Acc (%) on FF++ (LQ)			
	DeepFakes	Face2Face	FaceSwap	NeuralTextures	DeepFakes	Face2Face	FaceSwap	NeuralTextures
Microscopic path	99.05	98.57	98.39	90.05	91.26	89.72	90.01	73.4
Mesoscopic path	99.2	99.17	98.86	91.01	92.32	88.92	88.20	73.52
Macroscopic path	99.32	99.21	98.77	**91.52**	**94.4**	91.04	92.41	**75.09**
Multipath	**99.73**	**99.24**	**99.17**	91.30	93.83	**91.15**	**92.47**	74.41

Results in bold indicate the best results of the four ablation experiments.

Table 4: Ablation experiment on assigning weight to different semantics by adding attention module (ACC %).

	Acc (%) on FF++(HQ)				Acc (%) on FF++(LQ)			
	DeepFakes	Face2Face	FaceSwap	NeuralTextures	DeepFakes	Face2Face	FaceSwap	NeuralTextures
Without attention	**99.73**	99.09	**99.17**	91.2	93.83	91.15	**92.47**	74.41
With attention	99.49	**99.35**	99.06	**91.31**	**94.87**	**91.26**	91.13	**74.75**

Bold values indicate higher results in two experiments.

Table 5: Quantitative detection results of ACC (%) using FF++ dataset on high quality (C23 light compression) and low quality (C40 heavy compression) videos and AUC on TIMIT datasets. Bold font indicates the best result.

	Acc on FF++(HQ)	Acc on FF++(LQ)	AUC on TIMIT(HQ)	AUC on TIMIT(LQ)
Bayar and stamm [47]	88.68	61.6	86.50	88.33
InMesonet [4]	57.81	69.75	81.15	82.63
Rahmouni et al. [48]	—	58.10	—	—
Mesonet [4]	54.91	50.28	63.68	77.44
Zhou et al. [49]	—	—	73.5	83.5
Chollet [6]	91.87	72.93	93.64	88.24
Nirkin et al. [50]	—	75.00	—	—
Ours	**94.21**	**76.31**	**99.12**	**99.52**

Table 6: Face cross test results, using frame-level AUC (%) to compare our method with others on both benchmarks.

Method	FF-DF	Celeb-DFv2
Two-stream [49]	70.1	53.8
Meso4 [4]	84.7	54.8
MesoInception4 [4]	83.0	53.6
HeadPose [51]	47.3	54.6
FWA [52]	80.1	56.9
DSP-FWA [52]	93.0	64.0
VA-MLP [53]	66.4	55.0
VA-LogReg [53]	78.0	55.1
Xception [6]	93.65	64.52
Multitask [54]	76.3	54.3
Two branch [55]	93.18	**73.41**
Capsule [56]	96.6	57.5
Ours	**96.7**	66.7

Bold values indicate higher results in two experiments.

conducted experiments to verify the effectiveness of our proposed channel and spatial attention modules. It is valid for most datasets, as shown in Table 4. In particular, we find that NeuralTextures and Face2Face can have satisfactory effects in the most complex datasets of FaceForensics++.

Our overall accuracy on FaceForensics++ datasets exceeds that of many other previous methods, as shown in Table 5. Most of the work on the TIMIT dataset uses the AUC indicator. To evaluate the overall detection performance, we calculated the area under the curve (AUC), which is the area under the receiver operating characteristic (ROC) curve, whose maximum value is 1 and displays the results in Table 5. The AUC of our proposed method is higher than that of other methods, indicating better performance on compressed deepfake video detection.

4.5. Validation of Generalization on Celeb-DFv2. Cross-dataset validation was carried out to evaluate the generalization ability of the proposed method. The model was trained on FaceForensics++ and tested on Celeb-DFv2. We followed the setup of Celeb-DFv2 [10] to divide the test set and displayed the experimental index AUC scores in Table 6. It can be seen from the results that this method has a better generalization effect than most methods. Masi's [55] generalization on Celeb-DFv2 is better than ours, but the AUC score in the original dataset is far behind. Our approach has limitations, but it has always been a challenge to balance accuracy and generalization.

5. Conclusion

Although methods for deepfake detection of videos and images have made much progress, few methods consider multiple aspects of semantic information. This work proposes a new face forgery detection method, MSPNN, which can simultaneously capture micro, mesoscopic, and macro semantics to comprehensively distinguish forged images, with weights assigned automatically to the three semantics. The neural network can comprehensively capture different semantic information of an image. In view of the challenges of face tampering in a small-range, high-compression dataset, and cross-dataset, the proposed framework can effectively capture minor forged artifacts and macro forged traces, which can further improve the detection of high-compression forged images. This framework has good generalization as well. Furthermore, the proposed preprocessing method can improve the detection ability of our framework for low-quality counterfeit videos. Our future work will consider the combination of frequency domain information and brightness information at the separation point to integrate the corresponding features for deepfake detection.

Data Availability

The data supporting this work are from previously reported studies and datasets, which have been cited. The processed data are available at https://github.com/ondyari/FaceForensics/blob/master/dataset/README.md, https://conradsanderson.id.au/vidtimit/#downloads and https://github.com/yuezunli/celeb-deepfakeforensics.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this article.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (Nos. 62002313, 62101481, 62166047), Key Areas Research Program of Yunnan Province in China (No. 202001BB050076), Major Scientific and Technological Project of Yunnan Province (No. 202202AD080002), Yunnan Fundamental Research Projects of China (Nos. 202201AU070033, 202201AT070112), the fund project of Yunnan Province Education Department (No. 2022j0008), Key Laboratory in Software Engineering of Yunnan Province (No. 2020SE408), the open project of Engineering Research Center of Cyberspace in 2021-2022 (No. KJAQ202112012), and Industrial Internet Security Situation Awareness Platform of Yunnan Province.

References

[1] B. Chen, T. Li, and W. Ding, "Detecting deepfake videos based on spatiotemporal attention and convolutional lstm," *Information Sciences*, vol. 601, 2022.

[2] H. Li, W. Wang, C. Yu, and S. Zhang, "Swapinpaint: identity-specific face inpainting with identity swapping," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 7, 2021.

[3] Y. Luo, F. Ye, B. Weng, S. Du, and T. Huang, "A novel defensive strategy for facial manipulation detection combining bilateral filtering and joint adversarial training," *Security and Communication Networks*, vol. 2021, no. 7, Article ID 4280328, 10 pages, 2021.

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

[5] B. Han, X. Han, H. Zhang, J. Li, and X. Cao, "Fighting fake news: two stream network for deepfake detection via learnable srm," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 3, pp. 320–331, 2021.

[6] F. Chollet, "Xception: deep learning with depthwise separable convolutions," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1800–1807, Honolulu, HI, USA, July 2017.

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

[8] T. Huang, X. Zhang, W. Huang, L. Lin, and W. Su, "A multi-channel approach through fusion of audio for detecting video inter-frame forgery," *Computers & Security*, vol. 77, pp. 412–426, 2018, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167404818304243.

[9] P. Korshunov and S. Marcel, "Deepfakes: a new threat to face recognition? assessment and detection," *CoRR*, vol. abs/1812.08685, 2018, [Online]. Available: http://arxiv.org/abs/1812.08685.

[10] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-df: a large-scale challenging dataset for deepfake forensics," in *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3204–3213, Seattle, WA, USA, June 2020.

[11] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4396–4405, Long Beach, CA, USA, June 2019.

[12] F. Juefei-Xu, R. Wang, Y. Huang, Q. Guo, L. Ma, and Y. Liu, "Countering malicious deepfakes: survey, battleground, and horizon," *CoRR*, vol. abs/2103.00218, 2021, [Online]. Available: https://arxiv.org/abs/2103.00218.

[13] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein gan," 2017.

[14] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," 2018.

[15] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "Stargan: unified generative adversarial networks for multi-domain image-to-image translation," in *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8789–8797, Salt Lake City, UT, USA, June 2018.

[16] M. Liu, Y. Ding, M. Xia et al., "Stgan: a unified selective transfer network for arbitrary image attribute editing," in *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3668–3677, Long Beach, CA, USA, June 2019.

[17] "Zao app," 2019, [Online]. Available: https://zao-app.com.

[18] "Faceswap," 2016, [Online]. Available: https://github.com/deepfakes/faceswap.

[19] J. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *CoRR*, vol. abs/1703.10593, 2017, [Online]. Available: http://arxiv.org/abs/1703.10593.

[20] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman, "Synthesizing obama: learning lip sync from audio," *ACM Transactions on Graphics*, vol. 36, no. 4, pp. 1–13, 2017.

[21] Y. Yao, J. Ren, X. Xie, W. Liu, Y.-J. Liu, and J. Wang, "Attention-aware multi-stroke style transfer," in *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1467–1475, Long Beach, CA, USA, June 2019.

[22] Y. Jo, J. Park, and Sc-fegan, "Face editing generative adversarial network with user's sketch and color," in *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1745–1753, Seoul, Korea, November 2019.

[23] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang, "Second-order attention network for single image super-resolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, June 2019.

[24] B. Chen, W. Tan, G. Coatrieux, Y. Zheng, and Y.-Q. Shi, "A serial image copy-move forgery localization scheme with source/target distinguishment," *IEEE Transactions on Multimedia*, vol. 23, pp. 3506–3517, 2021.

[25] L. Nataraj, T. M. Mohammed, S. Chandrasekaran et al., "Detecting gan generated fake images using co-occurrence matrices," *Electronic Imaging*, vol. 2019, no. 5, pp. 532–537, 2019.

[26] J. Yang, A. Li, S. Xiao, W. Lu, X. Gao, and Mtd-net, "MTD-Net: learning to detect deepfakes images by multi-scale texture difference," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 4234–4245, 2021.

[27] J. Yang, S. Xiao, A. Li, W. Lu, X. Gao, and Y. Li, "Msta-net: forgery detection by generating manipulation trace based on multi-scale self-texture attention," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 7, pp. 4854–4866, 2022.

[28] H. Ling, J. Huang, C. Zhao, Y. Yao, J. Chen, and P. Li, "Learning diverse local patterns for deepfake detection with image-level supervision," in *Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7, Shenzhen, China, July 2021.

[29] J. Hu, S. Wang, and X. Li, "Improving the generalization ability of deepfake detection via disentangled representation learning," in *Proceedings of the 2021 IEEE International Conference on Image Processing (ICIP)*, pp. 3577–3581, Anchorage, AK, USA, September 2021.

[30] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan, "Recurrent convolutional strategies for face manipulation detection in videos," 2019.

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

[32] P. Chen, J. Liu, T. Liang et al., "Dlfmnet: end-to-end detection and localization of face manipulation using multi-domain features," in *Proceedings of the 2021 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, Shenzhen, China, July 2021.

[33] I. Huseynli and S. Varli, "Analyzing deep learning models' generalization ability under different augmentations on deepfake datasets," in *Proceedings of the 2021 6th International Conference on Computer Science and Engineering (UBMK)*, pp. 694–698, Ankara, Turkey, September 2021.

[34] G. Jia, M. Zheng, C. Hu et al., "Inconsistency-aware wavelet dual-branch network for face forgery detection," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 3, pp. 308–319, 2021.

[35] R. Durall, M. Keuper, and J. Keuper, "Watch your up-convolution: cnn based generative deep neural networks are failing to reproduce spectral distributions," in *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7887–7896, Seattle, WA, USA, June 2020.

[36] L. Li, C. Chen, L. Pan, J. Zhang, and Y. Xiang, "Sok: an overview of ppg's application in authentication," 2022, [Online]. Available: https://arxiv.org/abs/2201.11291.

[37] S. Makowski, P. Prasse, D. R. Reich, D. Krakowczyk, L. A. Jäger, and T. Scheffer, "Deepeyedentificationlive: oculomotoric biometric identification and presentation-attack detection using deep neural networks," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 4, pp. 506–518, 2021.

[38] H. Qi, Q. Guo, F. Juefei-Xu et al., "Deeprhythm: exposing deepfakes with attentional visual heartbeat rhythms," in *Proceedings of the 28th ACM International Conference on Multimedia, ser. MM '20*, pp. 4318–4327, Association for Computing Machinery, New York, NY, USA, 2020.

[39] X. Jin, D. Ye, and C. Chen, "Countering spoof: towards detecting deepfake with multidimensional biological signals," *Security and Communication Networks*, vol. 2021, no. 1, , Article ID 6626974, 8 pages, 2021.

[40] H. Jeon, Y. Bang, and S. S. Woo, "Fdftnet: facing off fake images using fake detection fine-tuning network," in *ICT Systems Security and Privacy Protection*, M. Hölbl, K. Rannenberg, and T. Welzer, Eds., pp. 416–430, Springer International Publishing, Cham, 2020.

[41] Z. Guo, G. Yang, J. Chen, and X. Sun, "Fake face detection via adaptive manipulation traces extraction network," *Computer Vision and Image Understanding*, vol. 204, 2021, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S107731422100014X, Article ID 103170.

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

[43] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826, Las Vegas, NV, USA, June 2016.

[44] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 8, pp. 2011–2023, 2020.

[45] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., pp. 499–515, Springer International Publishing, Cham, 2016.

[46] D. E. King, "Dlib-ml: a machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

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

'16, pp. 5–10, Association for Computing Machinery, New York, NY, USA, 2016.

[48] N. Rahmouni, V. Nozick, J. Yamagishi, and I. Echizen, "Distinguishing computer graphics from natural images using convolution neural networks," in *Proceedings of the 2017 IEEE Workshop on Information Forensics and Security (WIFS)*, pp. 1–6, Rennes, France, December 2017.

[49] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, "Two-stream neural networks for tampered face detection," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1831–1839, Honolulu, HI, USA, July 2017.

[50] Y. Nirkin, L. Wolf, Y. Keller, and T. Hassner, "Deepfake detection based on discrepancies between faces and their context," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 6111–6121, 2022.

[51] X. Yang, Y. Li, and S. Lyu, "Exposing deep fakes using inconsistent head poses," in *Proceedings of the ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8261–8265, Brighton, UK, May 2019.

[52] Y. Li and S. Lyu, "Exposing deepfake videos by detecting face warping artifacts," 2019.

[53] F. Matern, C. Riess, and M. Stamminger, "Exploiting visual artifacts to expose deepfakes and face manipulations," in *Proceedings of the 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pp. 83–92, Waikoloa Village, HA, USA, January 2019.

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

[55] I. Masi, A. Killekar, R. M. Mascarenhas, S. P. Gurudatt, and W. AbdAlmageed, "Two-branch recurrent network for isolating deepfakes in videos," in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds., pp. 667–684, Springer International Publishing, Cham, 2020.

[56] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Use of a capsule network to detect fake images and videos," *arXiv*, 2019.