systems

MDPI

Article

Multi-Feature Fusion Based Deepfake Face Forgery Video Detection

Zhimao Lai [1,2], Yufei Wang [3,*], Renhai Feng [4], Xianglei Hu [5] and Haifeng Xu [4]

[1] Immigration Management College (Guangzhou), China People's Police University, Guangzhou 510663, China; laizhimao@cppu.edu.cn
[2] Guangdong Provincial Key Laboratory of Information Security Technology, Guangzhou 510000, China
[3] China-Singapore International Joint Research Institute, Guangzhou 511356, China
[4] School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China; fengrenhai@tju.edu.cn (R.F.); xuhaifeng@tju.edu.cn (H.X.)
[5] School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510006, China; huxlei3@mail.sysu.edu.cn
* Correspondence: w.yf05@mail.scut.edu.cn

Abstract: With the rapid development of deep learning, generating realistic fake face videos is becoming easier. It is common to make fake news, network pornography, extortion and other related illegal events using deep forgery. In order to attenuate the harm of deep forgery face video, researchers proposed many detection methods based on the tampering traces introduced by deep forgery. However, these methods generally have poor cross-database detection performance. Therefore, this paper proposes a multi-feature fusion detection method to improve the generalization ability of the detector. This method combines feature information of face video in the spatial domain, frequency domain, Pattern of Local Gravitational Force (PLGF) and time domain and effectively reduces the average error rate of span detection while ensuring good detection effect in the library.

Keywords: multimedia forensics; feature fusion; deepfake; forgery detection; face swap

Citation: Lai, Z.; Wang, Y.; Feng, R.; Hu, X.; Xu, H. Multi-Feature Fusion Based Deepfake Face Forgery Video Detection. *Systems* **2022**, *10*, 31. https://doi.org/10.3390/systems10020031

Academic Editor: William T. Scherer

Received: 3 February 2022
Accepted: 5 March 2022
Published: 7 March 2022

1. Introduction

The human face usually provides important identity information; thus, many related studies were carried out, including face detection and recognition in 2D and 3D spaces [1–4]. In recent years, driven by computer graphics technology and deep learning, deep face forgery technology (Deepfake) achieved rapid development. It can replace the face of the target video character to the specified ones or let the target face repeat specific expression and action so as to realize the generation and replacement of high-fidelity faces [5]. Open-source tools and applications allowed ordinary users to change their faces according to their personal needs and generate high fidelity depth-forged face videos. The early application of deep forgery technology was only for the purpose of entertainment. However, due to the low technical threshold, high fidelity and strong deception of deep forgery face video, some criminals can easily forge face-changing videos of specific characters and for malicious use, which not only violates the privacy and reputation of the parties but also misleads public opinion, erodes social trust and even leads to serious political crisis. On the other hand, digital video evidence is an important class of electronic evidence in the judicial system, which is more and more widely used in various cases. It is very important to ensure its authenticity, which requires the support of digital video authenticity detection technology [6]. How to detect and defend against deep forgery face video has become one of the hottest issues concerned by governments, enterprises and individuals around the world.

In order to attenuate the harm caused by the deep face forgery technology, researchers carried out in-depth exploration face forgery video detection technology and put forward

the idea of detection from multiple perspectives such as the space domain, time domain and frequency domain. These methods achieved satisfactory detection performance on some data sets. However, they have defects such as low detection accuracy, poor generalization performance and weak anti-interference ability, which have great limitations and are difficult to be applied to more complex actual scenes. Based on such facts, this paper designs a feature fusion detection model that can extract features from the video spatial domain, frequency domain and time domain at the same time. The details are as follows: (a) spatial features of face images directly extract spatial features from face spatial images using the Xception network, (b) the frequency domain characteristics of face images are used to obtain the corresponding spectrum map by discrete Fourier transform of face images, and then extract it from the spectrum map through the Xception network, (c) the PLGF image is calculated, and then the PLGF feature is extracted by the Xception network and (d) time-domain features of face images are extracted by splicing and fusing the above three feature vectors of continuous multiple frames into the LSTM network structure. Finally, the output features of the LSTM network which fuse the information of face image spatial domain, frequency domain, PLGF and time domain are used for final classification detection. The source code of this work will be available in https://gitlab.com/test-2022/multi-feature-fusion-based-deepfake-detection (accessed on 3 March 2022).

2. Related Works

2.1. Generation Method of Deep Forged Face Video

Deep forgery is an image synthesis technology based on deep learning. It mainly uses a generative adversarial network, deep convolutional neural network and automatic encoder to forge a set of primitive faces and target faces as training data. One of the common applications of deep forgery is to replace one face in the video with another face, which is also known as face swapping [7]. The core idea of face swapping is to replace the face in the target video with the face in the source video and make the replaced face as realistic as possible through the corresponding detail processing so that the naked eye cannot distinguish whether the face in the output video is tampered. Because of changes involving identity attributes, face-changing techniques enable a specified person to appear in video scenes that never appeared before. The method of deep forgery of face video has been publicly implemented, which is mainly based on the structure of a denoising self-encoder and uses a supervised learning method to train a neural network for face replacement. In the following, Deep-Faceswap is taken as an example to introduce the generation process of deep forgery face video, and other deep forgery face generation methods also have similar generation processes [8].

Deep-Faceswap technology needs to use Dlib to extract the face in the source video and the face in the target video, then crop and align the extracted face and adjust the size to 64×64. In the training phase, primitive face A and target face B are used as training data to train a weight-sharing encoder for extracting the common facial attributes of A and B. In the decoding phase, A and B respectively train an independent decoder to learn the unique facial information of A and B and complete the corresponding face reconstruction. After the encoder and decoder of A and B are trained, in order to realize the face replacement between the original human face A and the target face B, the encoder is used to encode the facial attribute of B, and then the decoder of A is used to decode the facial attribute encoding feature of B to reconstruct the face. After that, the depth forgery face image with the appearance feature of face A and the facial expression action of face B is generated. The specific process is shown in Figure 1. Based on a similar idea, researchers proposed and developed more face replacement methods and achieved better replacement results.

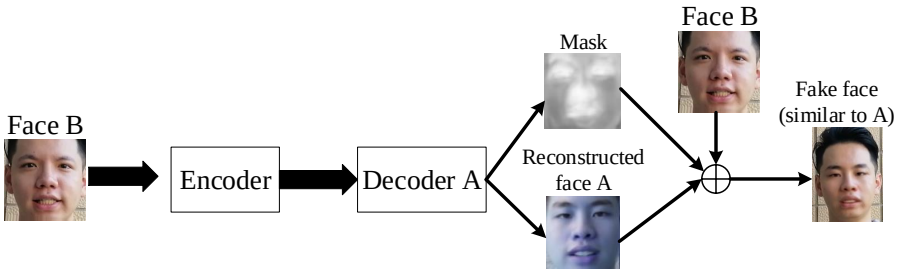

Figure 1. Deep-Faceswap fake face generation flow chart.

2.2. Deep Forged Face Video Database

In the process of deep forgery video detection, a database is indispensable, as it is mainly used to train and evaluate the performance of the detection model. There are four commonly used public face changing databases, namely DeepfakeTIMIT [9], Fake Face in the Wild (FFW) [10], FaceForensics++ [11] and DeepFakeDetection [12].

In DeepfakeTIMIT, face forgery videos are generated by the Swiss Idiap Institute using an open-source face replacement algorithm. The database selects 16 pairs of faces with similar facial features from VIDTIMIT to generate forgery videos, each video has 2 versions of different resolution sizes.

FFW is a face-changing database, released by the Biorecognition Laboratory of the Norwegian University of Science and Technology, which contains only face forgery video, using a variety of forgery techniques to generate forgery video.

In FaceForensics++, videos are collected from the YouTube video website, which contains 1000 real personage videos. A total of 4 face forgery techniques (DeepFake, Face2Face, FaceSwap and NeuralTextures) are used to generate 4 types of face forgery videos, and the number of each type of forgery video is 1000. In addition, FaceForensics++ database video also uses H.264 to compress the video into lossless video, high quality video and low quality video, and the corresponding compression rates are 0, 23 and 40, respectively, to simulate the compression of video in the actual transmission process.

Videos in DeepFakeDetection are jointly produced by Google and Jigsaw, which contains 363 original videos and 3068 counterfeit videos, with richer backgrounds and more diverse facial expressions. Similar to FaceForensics++, the DeepFakeDetection database also divides the video compression rate into C0, C23 and C40.

The algorithm presented in this paper will be tested on DeepfakeTIMIT, FaceForensics++ and DeepFakeDetection.

2.3. Deep Forgery Face Video Detection Method

The tamper detection methods of deep forgery face video are mostly based on the tampering traces introduced in the tampering process and the inconsistency of video frame images in spatial and temporal domains. The quality of forged face video production is related to the skin color difference and face action difference. When skin color difference is too large and the position of two face feature points cannot be accurately mapped, it is easy to cause obvious artifacts in the tampered video. Based on the tampering traces introduced in the process of depth forgery, researchers proposed corresponding detection methods to distinguish between fake face video images and real face video images. The proposed detection methods can be divided into two categories: machine learning method based on manual features and deep learning method. Zhang et al. [13] first proposed a classical machine learning method to detect face changing images. They first calculate speeded up robust features (SURF) descriptors, then generate bag of word (BOW) features by K-means method, obtain codebook histogram and input them into various classifiers, such as support vector machine (SVM), random forest (RF) and multi-layer perceptron, to distinguish face changing images and real face images by training. Due to the defects in the generation process of deeply forged faces, the stitching of the generated face region into the original image will introduce errors. Based on this, S. Lyu et al. [14] proposed a

3D attitude difference detection method based on the head posture position. This method uses the difference between the coordinate position of the central region of the face and the key points of 68 human faces extracted by Dlib as the features to distinguish true and false faces. The extracted features are standardized (mean and standard deviation) and then input into the SVM classifier to obtain the detection results. Matern et al. [15] summarized the artifacts left by current facial tampering and its processing, specifically the lack of consistent texture features in facial regions, such as missing reflection and detail in the eyes and teeth regions. Koopman et al. [16] proposed a method to classify true and false videos by using photo response non-uniformity (PRNU). PRNU is usually considered to be the camera fingerprint left by the camera in the image. Because face swapping will change the local PRNU mode of the face region in video frames, PRNU mode can be used as a feature to classify real and fake videos. By analyzing the images generated by GAN, the Horst Görtz Institute for IT-Security Research Team of the University of Bohongluer in Germany found that there were significant differences between the generated image and the real image in the frequency domain [17], which is caused by the up-sampling operation. Since the essence of GAN generating images is to transform low-dimensional noise vectors into high-dimensional images, the up-sampling operation cannot be avoided. Therefore, there must be grid characteristics in the frequency domain of the generated images. Habeeba [18] proposed a method using neural network to detect the visual artifacts of facial regions in non-natural images to distinguish true and false videos. Zhou et al. [19] proposed a dual-stream network to detect face tampering and considered the fusion of two features: face feature in spatial image and steganalysis feature of image block. Yu et al. [20] proposed a method for detecting face-changing images using separable convolutional neural networks, which combines the features after block with the whole image features to classify the images. Li et al. [21] used convolutional neural network (CNN) to extract the features of video frame images and then input the features of a specific number of continuous video frames into the recurrent neural network (RNN) to train the RNN to distinguish whether the video is face-changing video. Dolhansky [22] proposed three simple detection systems: (a) a small CNN model composed of six convolution layers and one fully connected layer to detect low-level image tampering, (b) the Xception network model using only face images for training and (c) the Xception network model using complete image training [23]. The existing deep-learning based methods show impressive performance. However, most of them just extract features from one or two domains. We try to fuse features extracted from more domains to improve the performance of detection.

3. Multi-Feature Fusion Based Deep Forgery Face Video Detection Method

3.1. Detection Framework

Compared with the real face, there is tampering trace information formed in the process of deep forgery. How to extract these tiny tampering traces is the key to distinguishing depth forgery faces. The intensity of deep forgery traces is too tiny to effectively detect and they have limited generalization ability based on a single feature. To this end, this paper proposes to detect from multiple perspectives such as space domain, time domain and frequency domain and designs a detection network that simultaneously extracts features from multiple feature spaces of face images, thus making the network have better generalization ability through multi-feature fusion.

This paper mainly considers the spatial domain features, frequency domain features, PLGF features and time domain features of face images. Spatial features of face images are extracted directly from face spatial images by the Xception network. The frequency domain features of face images need to obtain the corresponding spectrum map by discrete Fourier transform of face images and then extract it from the spectrum map through the Xception network. The PLGF feature of the face image needs to calculate the PLGF image of the face image first and then extract it from the PLGF image through the Xception network. Finally, the time domain features of the face image are extracted by merging three feature vectors of continuous multiple frames mentioned above into the LSTM network

structure. Finally, the output features of the LSTM network fuse the information of face image spatial domain, frequency domain, PLGF and time domain and will be used for final classification detection.

The detection framework of this paper is shown in Figure 2, in which the Xception network structure of feature extraction from the face image spatial domain, frequency domain and PLGF image is basically the same and finally output 2048-dimensional features. The three features compose 6144-dimensional features by splicing, which represent the spatial, frequency and PLGF fusion features of a face image. Time domain feature is extracted from fusion features of 10 face images through the double-layer LSTM network, which outputs 512-dimensional features [24]. Finally, binary classification results are output through a fully connected layer. The structure of the Xception network and the double-layer LSTM are determined by experiments. We have used Xception, ResNet50, InceptionV3, EfficientNet and DensNet201 as the feature extractor and found that the Xception network has the best performance among them. For the LSTM, we have used 1, 2 and 3 layer structures and found that the 2 layer structure has 2% to 3% higher accuracy than the 1 layer structure, while the 3 layer structure has almost the same performance with the 2 layer structure. Therefore, we selected the 2 layer LSTM for temporal feature extracting.

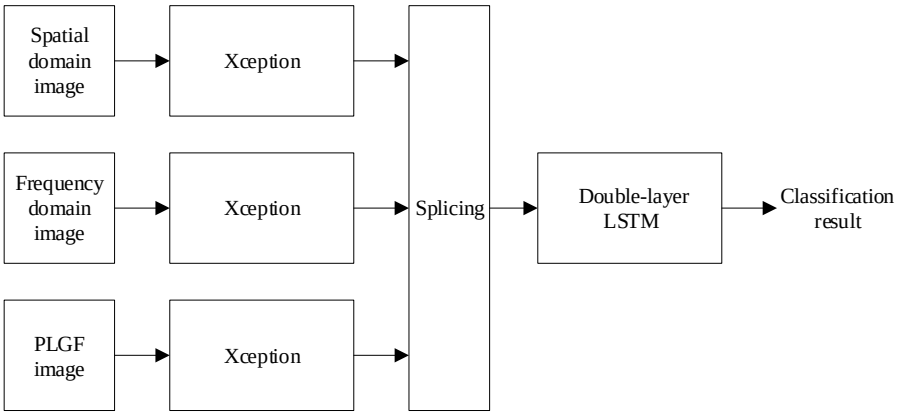

Figure 2. Detection framework.

3.2. Data Pre-Processing

The CNN model of the Dlib machine learning library in Python is used to detect the face region extracted from the video to be tested to obtain the face image. The face image is represented by I as an image with R, G, B three color channels and variable size. In order to keep the edge of the face in the extracted image, the bounding box of the face will be expended. The new bounding box has the same center with the old one, but its width and height are 1.3 times of the original ones.

For the input of the Xception network that extracts spatial characteristics, the size of I is adjusted to $224 \times 224 \times 3$ by bilinear interpolation and normalized. The obtained spatial image is denoted as I_S as the input of the network.

For the input of the Xception network to extract the frequency domain characteristics, DFT transform is performed on each channel in I, and the low frequency component is moved to the center to obtain the spectrum of each color channel. Assuming that the amplitude of the position of R channel (x, y) is $A^R(x, y)$, the value of the corresponding position of the frequency domain image is shown in Equation (1):

$$I_F^R(x, y) = \frac{20 \times \ln A^R(x, y)}{255} \tag{1}$$

The values at other positions are so deduced. Then, the size of each channel is adjusted to 224×224 by bilinear interpolation method, and the frequency domain input image I_F with size of $224 \times 224 \times 3$ is obtained.

For the Xception network input of PLGF image extraction, the horizontal gradient G_{hor} and vertical gradient G_{ver} are obtained by PLGF operator in the horizontal and vertical directions of the three color channels in I, respectively. PLGF convolution is expressed as follows:

$$\begin{gathered} G_{\mathrm{d}}[x,y] = \sum_{u=-1}^{1}\sum_{v=-1}^{1} f_{\mathrm{d}}[u,v]I[x-u,y-v], \mathrm{d} \in \{ \text{ hor }, \text{ ver } \} \\ f_{\text{hor}} = \begin{bmatrix} -1 & 0 & 1 \\ -\sqrt{2} & 0 & \sqrt{2} \\ -1 & 0 & 1 \end{bmatrix}, \quad f_{\text{ver}} = \begin{bmatrix} -1 & \sqrt{2} & -1 \\ 0 & 0 & 0 \\ 1 & \sqrt{2} & 1 \end{bmatrix} \end{gathered} \tag{2}$$

where f_{hor} and f_{ver} are 3×3 convolution kernels in the horizontal and vertical directions of the local gravity model (PLGF), respectively. $I[x,y]$ is the pixel value of coordinates (x,y), $G_{\mathrm{d}}[x,y]$ is the direction gradient of coordinates (x,y).

Then, according to the Lambert model, the horizontal and vertical gradients were separated by illumination to obtain the horizontal illumination separation gradient ISG_{hor} and the vertical illumination separation gradient ISG_{ver}. The operation of illumination separation is to divide the gradient and prevent its own pixel value of the minimum value of zero removal. Since light intensity changes slowly in a small area to a constant value L, the illumination component L can be eliminated to obtain the face texture features only related to the reflection coefficient. It has rich texture information and can be used as an effective feature for face detection authenticity. The specific expression of light separation is as follows:

$$ISG_{\mathrm{d}}[x,y] = \frac{G_{\mathrm{d}}[x,y]}{I[x,y]+\epsilon}, \mathrm{d} \in \text{hor}, \text{ver} \tag{3}$$

Then, the synthetic gradient ISG is obtained by linear activation of the horizontal and vertical light separation gradient, and the PLGF image is formed, as following:

$$ISG = arctan(\sqrt{(ISG_{\text{hor}})^2 + (ISG_{\text{ver}})^2}) \tag{4}$$

Finally, the PLGF image of each channel is bilinear interpolated, and its size is adjusted to 224×224, then the final PLGF input image I_P is obtained.

For the video to be detected, the down-sampling is carried out according to the frequency of each of the 5 frames, the actual frame image for detection is obtained so as to avoid the face image in the adjacent detection frame being too close, resulting in redundant information. The frame images obtained after down-sampling are processed according to the above method to obtain the corresponding I_S, I_F and I_P of each frame, that is, the corresponding data pre-processing is completed.

3.3. Xception Feature Extraction Network

Since the input I_S, I_F and I_P sizes as network inputs are completely consistent, the Xception network used to extract the spatial, frequency and PLGF features of face images also has the same structure. The Xception network structure used in this method is shown in Figure 3.

In Figure 3, Conv represents the normal convolution layer, SeparableConv represents the depth separable convolution layer, 3×3 and 1×1 represent the size of convolution kernel or pooling kernel, stride $= 2 \times 2$ represents the sliding step size of convolution kernel or pooling kernel is 2, and if it is not specially pointed out, the sliding step size is defaulted to 1.

3.4. Double-Layer LSTM Time Domain Feature Extraction Network

After extracting three 2048-dimensional features from spatial domain, frequency domain and PLGF through the above Xception network, the extracted features are spliced and fused to obtain 6144-dimensional features. Then, the 6144-dimensional fusion feature

of 10 frames of face images is input into the double-layer LSTM network structure, and the final 512-dimensional fusion feature is extracted, and the binary classification results of real face nuclear forgery face are output through the fully connected layer. The network structure of this part is shown in Figure 4.

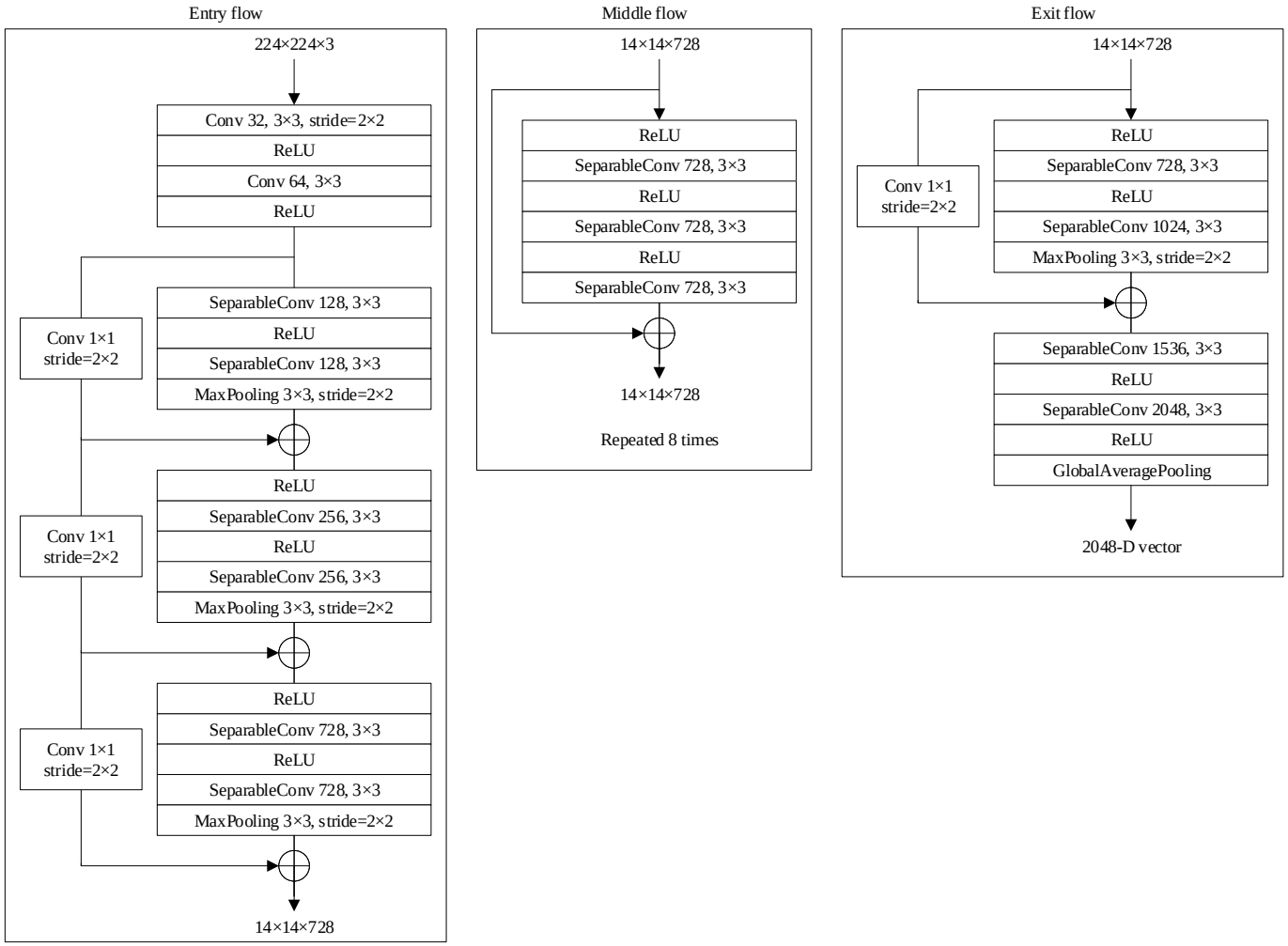

Figure 3. Xception network structure.

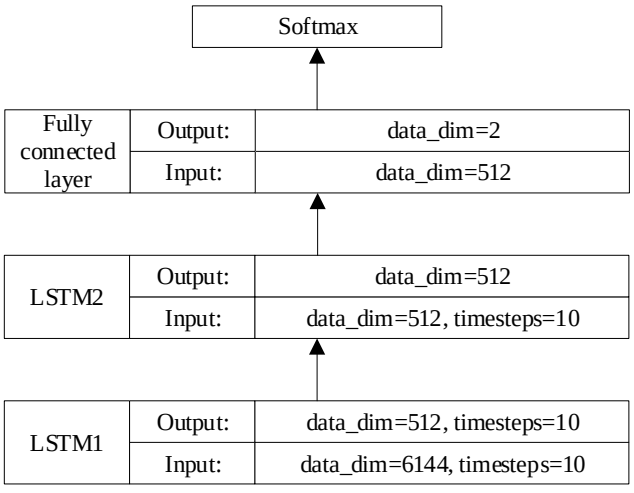

Figure 4. The double-layer LSTM time domain feature extraction network structure.

In Figure 4, the first layer input of LSTM is 6144-dimensional feature, and the output is 512-dimensional feature, which further integrates the features of various fields originally separated. The output of the first layer in LSTM contains 10 time steps, and the output features are sent to the second layer. The second layer in LSTM input is 512-dimensional features, and the output is 512-dimensional features with only one time step, which is the fusion of spatial domain, frequency domain, PLGF and time domain information of face images.

Finally, the 512-dimensional feature outputs a 2-dimensional vector through a fully connected layer and then outputs the binary classification results of real face or fake face contained in the video through the softmax activation function.

4. Experimental Results and Analysis

4.1. Introduction of Experimental Database

In order to evaluate the performance of the proposed method, we conduct experiments on three face forgery video databases: DeepFakeDetection (DFD), FaceForensics++ (FF++) and DeepfakeTIMIT (TIMIT). DFD database contains 1089 real videos and 9204 face forgery videos, which are divided into 3 different compression levels: synthetic compression rate 0 (C0), synthetic compression rate 23 (C23) and synthetic compression rate 40 (C40). The real video data come from 28 actors shooting in different scenes. The FF++ contains 1000 real videos and 4000 face forgery videos. There are 1000 face forgery videos synthesized by Deepfake tampering, which are divided into 3 different compression degrees: synthetic compression rate 0 (C0), synthetic compression rate 23 (C23) and synthetic compression rate 40 (C40). The real video data come from the video website YouTube. The TIMIT database contains 559 real videos and 640 face-changing videos. Face forgery videos include low quality (LQ) and high quality (HQ) videos.

4.2. Experimental Settings

In this experiment, a Nvidia GTX1080Ti card with 11 GB display memory is utilized. The operating system environment is Ubuntu 14.04, the programming language is Python 3.6 and the deep learning framework is Keras 2.2.5 based on TensorFlow 1.14.

During training, the database was divided into training set, verification set and test set according to the ratio of 7:2:1, and the batch size of the training sample was set to 32. The number of real and fake samples in the databases are usually unbalanced, so half the samples in each batch are randomly selected from real samples and the other samples are from fake samples. For the Xception network, the Adam method is used to optimize, and the learning rate is set to 0.0001. For the LSTM network, RMSProp method is used to optimize, and the learning rate is set to 0.001. The automatic decline strategy of learning rate is adopted in training. If the loss does not decline after five iterations, the learning rate is set to 0.5 times that of the original. If the loss does not decrease after 10 iterations, the network model is considered to have converged, and then the training is terminated. It takes about 15 s for one batch during training.

Half total error rate (HTER) is used as the evaluation index, which is the average value of false alarm rate and missed detection rate of the algorithm under the decision threshold, as defined in Equation (5). Among them, False Acceptance Rate (FAR) refers to the error acceptance rate, that is, the ratio of the algorithm to judge the face changing face as the real face. False Rejection Rate (FRR) refers to the error rejection rate, that is, the algorithm to judge the real face as the ratio of the face changing face, defined as Equations (6) and (7), where N_{f2t} refers to the number of times that the face changing face is judged as the real face, N_f refers to the total number of attacks on the face changing face, N_{t2f} refers to the number of times that the real face is judged as the face changing face. N_t refers to the total number of real face detection. The lower the HTER, the better the performance of the algorithm.

$$HTER = \frac{FAR + FRR}{2} \tag{5}$$

$$FAR = \frac{N_{f2t}}{N_f} \tag{6}$$

$$FRR = \frac{N_{f2t}}{N_t} \tag{7}$$

4.3. Ablation Experiment

In order to verify the effect of each branch structure of the algorithm, the corresponding ablation experiments are carried out in this section. The model was trained on the DFD (C23) database and then tested on the DFD (C23) database, FF++ (C0) database, FF++ (C23) database and TIMIT database, and HTER was selected as the evaluation index of the algorithm. The experimental results are shown in Table 1.

Table 1. The HTER of ablation experiment (%).

Training Data Set	**DFD (C23)**			
Test Data Set	**DFD (C23)**	**FF++ (C0)**	**FF++ (C0)**	**TIMIT**
Only spatial	4.68	17.69	22.87	20.18
Only frequency	10.47	24.17	27.88	27.05
Only PLGF	6.85	19.26	24.69	22.19
Without LSTM	3.57	16.18	21.17	17.69
Complete method	2.91	14.17	18.32	15.34

It can be seen from Table 1 that the complete method proposed in this paper has the best performance in both in-library and cross-library tests, followed by the network structure effect of directly integrating spatial, frequency and PLGF characteristics without using the double-layer LSTM network. The effect of using three branches alone is not as good as that of fusing features. The effect of using airspace image alone is better than that of using the other two images alone, while the performance of using frequency domain image alone is the lowest. In summary, each branch structure of the multi-feature fusion method proposed in this paper plays a role in improving the performance of the method.

4.4. Comparing with Other Algorithms

In order to further verify the performance of the proposed algorithm, this section compares the proposed method with the depth-forged face detection method published in recent years and trains them on the DFD (C23) and the FF++ (C0 and C23), respectively. Then, tests are conducted on the DFD (C23), the FF++ (C0), the FF++ (C23) and the TIMIT, and HTER is selected as the algorithm evaluation index. The experimental results are shown in Tables 2 and 3.

From Tables 2 and 3, it can be seen that the proposed algorithm has the best performance when DFD (C23) is used as the training sample, in-library detection and most cross-library detection; only when FF++ (C0) is used as the test sample, the performance is not as good as MISLnet. While using FF++ (C0 and C23) as training samples, the proposed algorithm is not the best but also has considerable performance, showing the robustness and generalization ability of the algorithm.

Table 2. The HTER comparison of DFD (C23) trained model (%).

Training Data Set	**DFD (C23)**			
Test Data Set	**DFD (C23)**	**FF++ (C0)**	**FF++ (C23)**	**TIMIT**
MesoInception [25]	7.26	21.54	25.76	35.88
MISLnet [26]	3.26	5.75	16.21	18.73
ShallowNet [27]	5.84	19.34	21.44	27.45
Xception [23]	4.38	18.84	22.31	20.60
S-MIL-Vb [28]	3.66	22.51	21.88	30.56
S-MIL-Fb [28]	6.29	25.39	26.34	34.43
FFD-Vgg-16 [29]	3.22	19.63	24.19	34.86
Proposed algorithm	2.91	14.17	18.32	15.34

Table 3. The HTER comparison of F++ (C0 and C23) trained model (%).

Training Data Set	**F++ (C0 and C23)**			
Test Data Set	**DFD (C23)**	**FF++ (C0)**	**FF++ (C23)**	**TIMIT**
MesoInception [25]	3.08	8.14	28.11	22.29
MISLnet [26]	0.72	1.98	25.06	24.26
ShallowNet [27]	1.76	4.37	28.27	25.55
Xception [23]	0.95	1.88	26.61	21.46
S-MIL-Vb [28]	1.09	2.58	17.23	13.14
S-MIL-Fb [28]	2.85	4.03	12.75	33.84
FFD-Vgg-16 [29]	1.25	2.77	27.48	28.97
Proposed algorithm	1.24	2.25	24.34	25.83

5. Conclusions

Aiming at the defects of low detection accuracy, poor generalization performance and weak anti-interference ability of existing deep network face-changing video tampering detection algorithms, this paper proposes a deep forgery face video detection method based on multi-feature fusion. The spatial domain, frequency domain and PLGF feature information of face images are extracted by the Xception network, and the time domain feature information of face images is extracted by the double-layer LSTM network. The experimental results show that the proposed method has good in-library and cross-library detection performance and strong generalization ability.

Author Contributions: Conceptualization, Z.L. and Y.W.; methodology, Y.W.; software,Y.W.; validation, X.H., R.F. and H.X.; formal analysis, R.F.; investigation, H.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Young Teachers Program of Scientific Innovation Project from China People's Police University (ZQN2020028), the Science and Technology Project of Hebei Education Department (QN2021417), the Opening Project of Guangdong Province Key Laboratory of Information Security Technology (Grant No. 2020B1212060078-05), the Key Research Project from China People's Police University (2019zdgg012), the Research Project on Social Science Development of Hebei Province in 2021 (No. 20210101014), the China-Singapore International Joint Research Institute (Grant No. 206-A018001), the Foundation of Shenzhen City (Grant No. JCYJ20160422093217170), the Natural Science Foundation of China (Grant No. 61601309), and the State Grid Electric Power Co., Ltd. (Grant No. 5700-202025165A-0-0-00).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

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

References

1. Marcolin, F.; Vezzetti, E.; Monaci, M.G. Face perception foundations for pattern recognition algorithms. *Neurocomputing* **2021**, *443*, 302–319. [CrossRef]
2. Payal, P.; Goyani, M.M. A comprehensive study on face recognition: Methods and challenges. *Imaging Sci. J.* **2020**, *68*, 114–127. [CrossRef]
3. Ulrich, L.; Vezzetti, E.; Moos, S.; Marcolin, F. Analysis of RGB-D camera technologies for supporting different facial usage scenarios. *Multimed. Tools Appl.* **2020**, *79*, 29375–29398. [CrossRef]
4. Keck, M.; Davis, J.W. Recovery and reasoning about occlusions in 3D using few cameras with applications to 3D tracking. *Int. J. Comput. Vis.* **2011**, *95*, 240–264. [CrossRef]
5. Tolosana, R.; Vera-Rodriguez, R.; Fierrez, J.; Morales, A.; Ortega-Garcia, J. Deepfakes and beyond: A survey of face manipulation and fake detection. *Inf. Fusion* **2020**, *64*, 131–148. [CrossRef]
6. Katarya, R.; Lal, A. A Study on Combating Emerging Threat of Deepfake Weaponization. In Proceedings of the 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), Palladam, India, 7–9 October 2020; pp. 485–490.
7. FaceSwap. Available online: https://github.com/MarekKowalski/FaceSwap (accessed on 30 January 2022).
8. faceswap. Available online: https://github.com/deepfakes/faceswap (accessed on 30 January 2022).
9. Korshunov, P.; Marcel, S. Vulnerability assessment and detection of deepfake videos. In Proceedings of the 2019 International Conference on Biometrics (ICB), Crete, Greece, 4–7 June 2019; pp. 1–6.
10. Khodabakhsh, A.; Ramachandra, R.; Raja, K.; Wasnik, P.; Busch, C. Fake face detection methods: Can they be generalized? In Proceedings of the 2018 International Conference of the Biometrics Special Interest Group (BIOSIG), Darmstadt, Germany, 26–28 September 2018; pp. 1–6.
11. Rössler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; Nießner, M. FaceForensics++: Learning to detect manipulated facial images. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 1–11.
12. Contributing Data to Deepfake Detection Research. Available online: https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html (accessed on 30 January 2022).
13. Zhang, Y.; Zheng, L.; Thing, V.L.L. Automated face swapping and its detection. In Proceedings of the 2017 IEEE 2nd International Conference on Signal and Image Processing (ICSIP), Singapore, 4–6 August 2017; pp. 15–19.
14. Yang, X.; Li, Y.; Lyu, S. Exposing deep fakes using inconsistent head poses. In Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 8261–8265.
15. Matern, F.; Riess, C.; Stamminger, M. Exploiting visual artifacts to expose deepfakes and face manipulations. In Proceedings of the 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW), Waikoloa Village, HI, USA, 7–11 January 2019; pp. 83–92.
16. Koopman, M.; Rodriguez, A.M.; Geradts, Z. Detection of deepfake video manipulation. In Proceedings of the 20th Irish Machine Vision and Image Processing Conference (IMVIP), Belfast, Ireland, 29–31 August 2018; pp. 133–136.
17. Frank, J.; Eisenhofer, T.; Schönherr, L.; Fischer, A.; Kolossa, D.; Holz, T. Leveraging Frequency Analysis for Deep Fake Image Recognition. *arXiv* **2020**, arXiv:2003.08685.
18. Habeeba, M.A.S.; Lijiya, A.; Chacko, A.M. Detection of Deepfakes Using Visual Artifacts and Neural Network Classifier. In *Innovations in Electrical and Electronic Engineering*; Springer: Singapore, 2021; pp. 411–422.
19. Zhou, P.; Han, X.; Morariu, V.I.; Davis, L.S. Two-stream neural networks for tampered face detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 21–26 July 2017; pp. 1831–1839.
20. Yu, C.M.; Chang, C.T.; Ti, Y.W. Detecting Deepfake-forged contents with separable convolutional neural network and image segmentation. *arXiv* **2019**, arXiv:1912.12184.
21. Li, X.; Yu, K.; Ji, S.; Wang, Y.; Wu, C.; Xue, H. Fighting against deepfake: Patch and pair convolutional neural networks (PPCNN). In Proceedings of the Web Conference 2020, Taipei, Taiwan, 20–24 April 2020; pp. 88–89.
22. Dolhansky, B.; Howes, R.; Pflaum, B.; Baram, N.; Ferrer, C.C. The deepfake detection challenge (dfdc) preview dataset. *arXiv* **2019**, arXiv:1910.08854.
23. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258.
24. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
25. Afchar, D.; Nozick, V.; Yamagishi, J.; Echizen, I. Mesonet: A compact facial video forgery detection network. In Proceedings of the 2018 IEEE International Workshop on Information Forensics and Security (WIFS), Hong Kong, China, 11–13 December 2018; pp. 1–7.
26. Bayar, B.; Stamm, M.C. Constrained convolutional neural networks: A new approach towards general purpose image manipulation detection. *Trans. Inf. Forensics Secur.* **2018**, *13*, 2691–2706. [CrossRef]
27. Tariq, S.; Lee, S.; Kim, H.; Shin, Y.; Woo, S.S. Detecting both machine and human created fake face images in the wild. In Proceedings of the 2nd International Workshop on Multimedia Privacy and Security, Toronto, ON, Canada, 15 October 2018; pp. 81–87.

28. Li, X.; Lang, Y.; Chen, Y.; Mao, X.; He, Y.; Wang, S.; Xue, H.; Lu, Q. Sharp Multiple Instance Learning for DeepFake Video Detection. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 1864–1872.
29. Dang, H.; Liu, F.; Stehouwer, J.; Liu, X.; Jain, A.K. On the detection of digital face manipulation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 5781–5790.