Hindawi
Security and Communication Networks
Volume 2021, Article ID 6626974, 8 pages
https://doi.org/10.1155/2021/6626974

WILEY | Hindawi

Research Article

Countering Spoof: Towards Detecting Deepfake with Multidimensional Biological Signals

Xinlei Jin, Dengpan Ye, and Chuanxi Chen

School of Cyber Science and Engineering, Wuhan University, Wuhan 430072, China

Correspondence should be addressed to Dengpan Ye; yedp@whu.edu.cn

Received 27 December 2020; Revised 20 March 2021; Accepted 10 April 2021; Published 22 April 2021

Academic Editor: Beijing Chen

The deepfake technology is conveniently abused with the low technology threshold, which may bring the huge social security risks. As GAN-based synthesis technology is becoming stronger, various methods are difficult to classify the fake content effectively. However, although the fake content generated by GANs can deceive the human eyes, it ignores the biological signals hidden in the face video. In this paper, we proposed a novel video forensics method with multidimensional biological signals, which extracting the difference of the biological signal between real and fake videos from three dimensions. The experimental results show that our method achieves 98% accuracy on the main public dataset. Compared with other technologies, the proposed method only extracts fake video information and is not limited to a specific generation method, so it is not affected by synthetic methods and has good adaptability.

1. Introduction

With the rapid advancement of computer vision and digital content processing technology, face tampering is no longer limited to pictures, some deep learning technologies (e.g., deepfake) can be utilized to generate human faces in videos, which are very similar to natural face videos taken by using digital cameras, but it is difficult to distinguish them with the naked eyes. The recent research study by Korshunov [1] shows that fake videos can easily deceive the face recognition system, and some serious security risks, such as fake news, have been raised by them.

Deepfake technology is the result of scientific and technological progress and the rapid development of artificial intelligence technology, and it has broad application prospects. For example, deepfake technology is used in entertainment industries such as films, which can save time and labor costs. However, if this technology is abused by criminals, it will also cause a serious crisis, and it can even forge the speeches of world leaders, seriously endangering political security. Therefore, the forensics of deepfake videos is of great significance. At present, the forensics method of deepfake video is mainly based on intraframe or interframe information by analyzing the difference between real and fake videos.

In this paper, we propose a deepfake video forensics method based on multidimensional biological signals. Recent work shows that heart rate signals can be used to effectively distinguish between real and fake videos [2, 3]. Although GANs can generate fake content that deceive human eyes, it destroys the original biological signals of the real video, such as heart rate signals. Therefore, we can classify the real and fake videos by extracting and analyzing the biological signals in the videos. Our main contributions are as follows:

(1) We propose a synthetic video forensics method, which mainly analyses the different biological signals between real and fake videos to detect the spoofed content.

(2) We further explore the distinct information in the multidimensional scene to ensure the technological efficiency. That is, we utilize the RGB space to concentrate on the color variations, the YUV space to concentrate on brightness alteration, and the chrominance method to reduce noise effects.

(3) We analyzed the shortcomings of traditional photoplethysmography (PPG) and used a deep neural network to realize the classification of real and fake videos. The experimental results show that the deep models can reach high detection accuracy, which is about 98% on the main public dataset.

The rest of this paper is organized as follows. Section 2 introduces related work, including the development of PPG and deepfake video forensics. Section 3 describes the proposed method in detail. Section 4 shows the details and results of our experiment. In Section 5, we conclude and give the future work.

2. Related Work

2.1. Deepfake. Deepfakes are fake videos digitally manipulated to depict people saying and doing things that never actually happened. Deepfakes rely on neural networks that analyze large sets of data samples to learn to mimic a person's facial expressions and mannerisms. The process involves feeding footage of two people into a deep learning algorithm to train it to swap faces.

The overall pipeline of the basic deepfake is shown in Figure 1. The autoencoder is usually formed by two convolutional neural networks (the encoder and the decoder). The encoder converts the input target's face to a vector. There is only one single encoder regardless of the identities of the subjects to ensure the encoder captures identity-independent attributes such as facial expressions. On the other hand, each identity has a dedicated decoder, which generates a face of the corresponding subject from the vector. Specifically, an encoder-decoder pair is formed alternatively using encoder and decoder for input face of each subject, and their parameters are optimized to minimize the reconstruction errors. The parameter update is performed with the back-propagation until convergence. The training stage can be stated as

$$\begin{aligned} \min L_A &= \frac{1}{N}\sum_{i=1}\left\|F_i - D_A\left(E\left(F_i;\,\theta\right);\,\phi_A\right)\right\|^2, \\ \min L_B &= \frac{1}{N}\sum_{i=1}\left\|F_i - D_B\left(E\left(F_i;\theta\right);\phi_B\right)\right\|^2, \end{aligned} \tag{1}$$

where L denotes the loss value of the autoencoder; N is the number of input data of the network; F_i is the input face image; θ is the weight of encoder E; and Φ is the weights of decoder D.

In the converting stage, the trained decoder B is used to decode the latent vector of face A to obtain the face-swapping image of A. Similarly, we can use the trained decoder A to decode the latent vector of face B to obtain the face-swapping image of B. The converting stage can be stated as

$$\begin{aligned} F_A' &= D_B\left(E\left(F_A;\,\theta\right);\,\phi_B\right), \\ F_B' &= D_A\left(E\left(F_B;\theta\right);\phi_A\right), \end{aligned} \tag{2}$$

where F denotes the original face and F' denotes the fake face.

2.2. Biological Signals. Biological signal extraction was originally used in the medical field to detect whether the patient's heart rate (HR) or other signals are normal, so that the doctor can observe the abnormal biological signal of the patient in time. However, electrocardiogram (ECG) leads, pulse oximeters, and other detectors all require specific sensors to be connected to the human body. To avoid the use of intrusive sensors, computer vision researchers have proposed a method of noncontact remote HR measurements, based on observing subtle changes in color and motion in the RGB video, such as remote photoplethysmography (PPG) [4, 5].

Balakrishnan et al. [6] show that heart activity can cause head movements, which can be used to extract heart rate estimates from video streams. Tulyakov proposed a chrominance method, which can effectively improve the accuracy of heart rate estimation [5]. Niu proposed a remote heart rate estimation method based on deep learning and achieved good results [7].

2.3. Forgery Detection. To deal with the possible harm caused by deepfake videos, researchers are exploring effective methods to classify real and fake videos. Because deepfake is also a forgery of images, early detection methods can learn from the forgery detection method of images. Recently, a bunch of high-efficient detectors with the new algorithms have been proposed to improve the performance of tampering detection and localization [8, 9]. Also, in order to specifically detect deepfake forgery, researchers classify real and fake videos based on intraframe information, interframe information, or special artifact.

Nguyen et al. [10] proposed a capsule network that can detect various kinds of attacks, from presentation attacks using printed images and replayed videos to attacks using fake videos created using deep learning. It uses fewer parameters than traditional convolutional neural networks with similar performance. Do et al. [11] used a deep convolutional neural network (VGGFace) for detecting real/fake images from GANs. Afchar et al. [12] exploited features at a mesoscopic level, instead of purely microscopic and macroscopic features, and proposed mesonet and meso-4 net, which have a low number of parameters. Bonettini et al. [13] combined CNNs, attention layers, and siamese training and achieved good performance on DFDC. Li and Lyu [14] created negative data only using a simple image processing operation, rather than using deepfake to produce, and then used CNN models to classify the videos. Zhao [15] formulated deepfake detection as a fine-grained classification problem and proposed a new multiattentional deepfake detection network. Liu [16] proposed a novel Spatial-Phase Shallow Learning (SPSL) method, which combines the spatial image and phase spectrum to capture the upsampling artifacts of face forgery to improve the transferability.

Güera and Delp [17], based on temporal inconsistencies between frames, used CNN (frame feature extraction) and RNN (temporal sequence analysis) for real and fake video

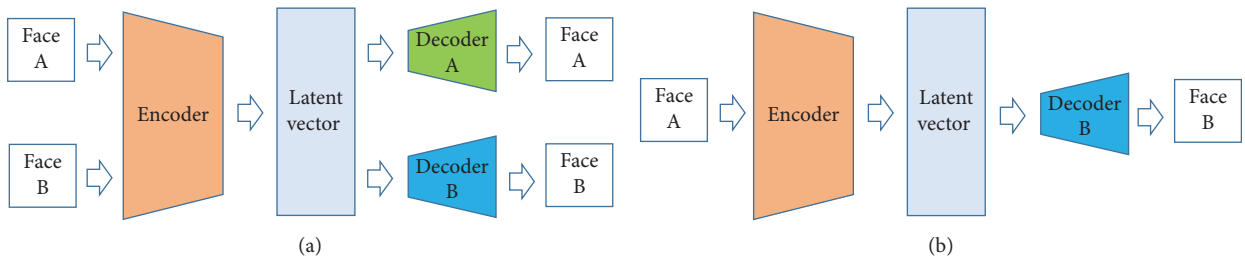

Figure 1: Overview of the deepfake procedure. (a) The training stage of deepfake. (b) The converting stage of deepfake.

classification. Sabir et al. [18] also proposed the CNN + RNN method, but they used face alignment and bidirectional recurrency.

Agarwal et al. [19] tracked facial and head movements and then extracted the presence and strength of specific action units and classified real and fake video by SVM. Li et al. [20] used CNN and RNN to detect abnormal blinking in fake videos. Yang et al. [21] classified real and fake videos by the inconsistency of 3D head poses. Li et al. [22] detected whether the input image can be decomposed into the blending of two images from different sources. Wang et al. based on monitoring neuron behaviors to spot AI-synthesized fake faces [23].

3. Method

In this section, we first analyze the discrepant biological signals between real and fake videos. Then, we point out the inefficiency of the traditional PPG method for detecting the deepfake video. Lastly, we propose a deepfake video forensics method based on the inconsistency of biological signals, and experimental results of the evaluation verify the effectiveness of our method.

3.1. Deepfake Detection with Biological Signals. Although PPG technology has been developed for a long time, it is not easy to extract heart rate signals in an unrestricted environment. We analyzed the method of manually extracting heart rate signals from the face video using computer vision; Figure 2 shows that these methods cannot distinguish fake videos from real videos. We selected a pair of real and fake videos from the DeepFakeDetection (DFD) dataset and used the Kalman filter [24] method to estimate heart rate signals from them. The result shows that the difference in heart rate signals between real and fake videos is not obvious.

Generally, to eliminate motion artifacts and noise caused by environmental changes and extract pure heart rate signals better, the videos are always processed by denoising and filtering. However, these technologies destroy the abnormal heart rate signals in the fake video, which cause the weak classification effect. Therefore, we map the video to ppg_map and classify it through the deep network to achieve the effect of deepfake video classification based on different heart rate extraction algorithms. In detail, given a video Vmmc5 (= {T1, T2, T3 ..... Tk}) including k frames, for each frame, we first extract the face and make face alignment. Then, the skin segmentation is performed to remove the influence of the background. Next, the face image is divided into n blocks (R1, R2, R3 ... Rn), which are independent on each other. Lastly, we calculate the signal value in each block from multidimension. The signal values of different blocks in the same frame are arranged in columns, and the signal values of the same block in different frames are arranged in rows to form our ppg_map. Then, these ppg_maps are used to train the CNN classification model, as shown in Figure 3.

In the process of generating ppg_map, it is necessary to avoid the adverse effects of the head movement and background of the characters. We will discuss this in detail in Section 3.2.

3.2. ppg_Map Generation. The beating of the human heart causes the periodic constriction of blood vessels, which affects the skin's reflection of light. This change is not easily detectable by the human eyes, but it can be detected and recorded by optical instruments. The facial area in the face video can well reflect the heart rate information of the human body. So, we located the facial area and extracted the biological signals.

3.2.1. Face Detection and Alignment. In order to make the detection faster and simpler, the Viola and Jones method [25] is utilized to detect human faces. However, because the faces in the video will not be fixed at a certain position and angle, we align the detected faces by rotating the face to keep both eyes at the same level. On the other hand, the face area detected by the Viola–Jones method is larger than the real face area and contains more background area; we further adjust the region of interest (ROI). In other words, we located 81 landmarks and used four points (1, 8, 15, and 71) as reference points to adjust the face region (Figure 4) to make the ROI include as many face regions as possible.

3.2.2. Skin Detection and Segmentation. The biological signals are extracted from the facial skin, so we reduce the negative influence of other nonskin areas, such as eyes, hair, and background areas. Meanwhile, this will also reduce the disturbance caused by eye blinking and lip motions. Consequently, in the video frame, we first adopt the skin detection algorithm to gain the main facial skin information. Then, as a mask, the skin area is used to extract the facial skin and remove the background and nonskin areas.

3.2.3. Blocks Division and Signal Extraction. Now, we have made the skin detection and segmentation to make

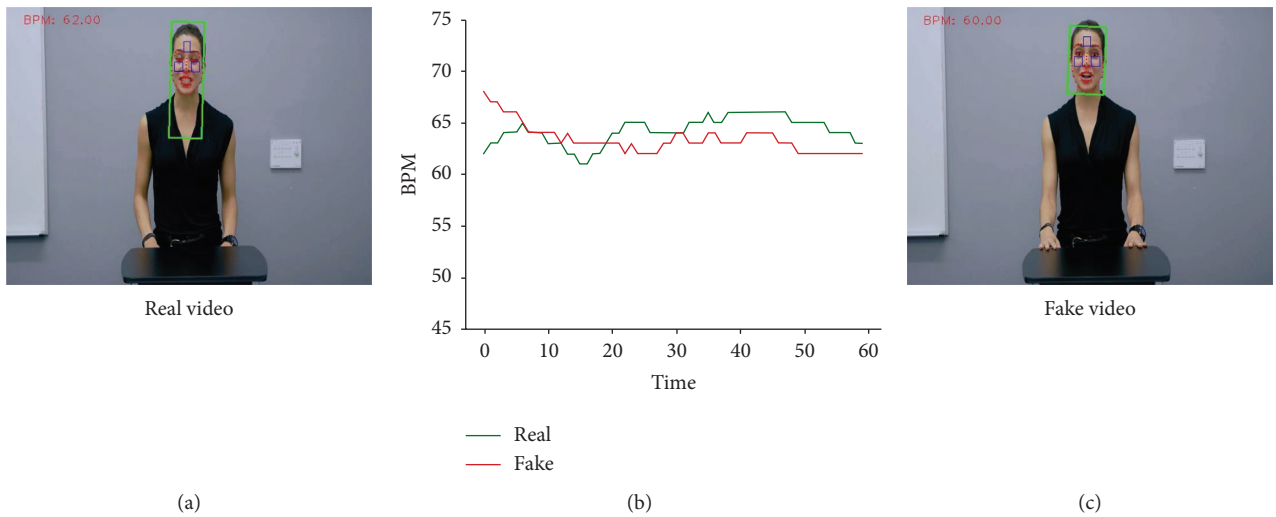

Figure 2: Comparison of real and fake video heart rate. The horizontal axis represents the number of frames of the video, and the vertical axis represents the detected heart rate.

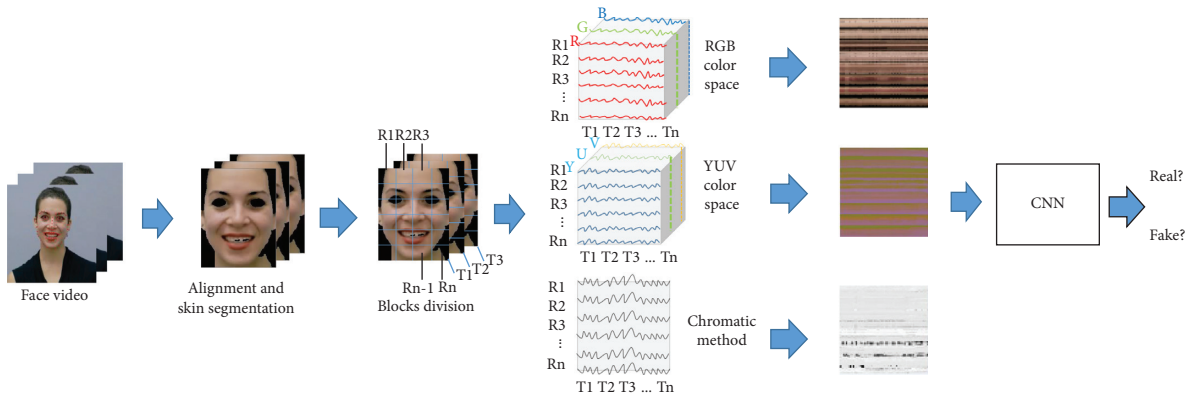

Figure 3: Overview of the proposed method.

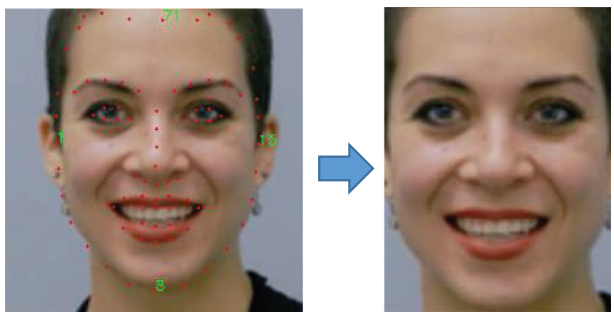

Figure 4: Refining the face ROI. We locate 81 landmarks on the face and use the four points (1, 8, 15, and 71) as benchmarks to further adjust the detected facial area to reduce the background area.

biological signals clearer. Then, the video frame is divided into m ∗ n blocks for extracting biological signals from every block. The PPG method mainly extracts heart rate signals from three dimensions [4–6]. That is, the RGB dimension intuitively reflects the changes in the color of the human face, the YUV dimension pays more attention to changes in brightness, and the chrominance dimension can effectively eliminate environmental artifacts and errors caused by head movement. So, we extract biological signals from RGB color space, YUV color space, and chrominance dimension.

(1) Color Space Dimension. We split the block into three channels of RGB (YUV) and calculate the pixel average of each channel for all blocks. Then, 3 sequences of length m ∗ n can be derived in a frame. Meanwhile, the same block in different frames is also changed with frames. When these procedures are employed in T frames, we can get a three-dimensional matrix with the shape $T * N * 3$, where T denotes the number of frames, N denotes the number of blocks, and 3 represents three channels (RGB or YUV). Each row of the matrix represents the change of the same block on different frames, and each column represents the changes of different blocks in the same frame.

(2) Chrominance Dimension. We calculate the average chrominance of each block [5]. For each pixel, the chrominance signal C is computed as the linear combination of two signals Xf and Yf:

$$C = \mathrm{Xf} - \alpha \mathrm{Yf},$$
$$\alpha = \frac{\sigma(\mathrm{Xf})}{\sigma(\mathrm{Yf})}, \tag{3}$$

where σ (Xf) and σ (Yf) denote the standard deviations of Xf and Yf. The signals Xf and Yf are band-passed filtered signals obtained, respectively, from the signals X and Y,

$$X = 3\mathrm{Rn} - 2\mathrm{Gn},$$
$$Y = 1.5\mathrm{Rn} + \mathrm{Gn} - 1.5\mathrm{Bn}, \tag{4}$$

where Rn, Gn, and Bn are the normalized values of the individual color channels. When we adopt the operations for all blocks and T frames, we can get a two-dimensional matrix with the shape $T * N$.

These matrices are stored as color maps (three-dimensional matrix) and grayscale maps (two-dimensional matrix) to form the corresponding ppg_map. Then, we move the sliding window to generate the next ppg_map the same way.

3.3. CNN-Based Classification. We use a CNN classifier to classify the generated ppg_map. The network consists of six convolutional layers, using the 'relu' activation function, followed by a flatten layer. There are two fully connected layers after convolutional layers. The last fully connected layer uses 'softmax' as the activation function and outputs the scores of the positive and negative classes. In order to avoid overfitting, we added a dropout layer, as shown in Figure 5.

For each dimension in Section 3.2.3, we trained the model and get the accuracy on the testing set. Furthermore, we combine the signals of three dimensions to make the final decision.

4. Results

In this section, we will introduce the details of our experiment. First, we describe the dataset we used. Then, we provide detailed experimental settings and the result of the experiment.

4.1. Dataset. We used three public datasets to train and test our method. For each dataset, we generated the ppg_maps and divided it into a training set, validation set, and testing set according to the ratio of 6 : 2:2. We optimize our model on the training set and validation set and then get the forensics accuracy on the testing set.

4.1.1. Face Forensics++. The FF++ dataset is proposed by Andreas [26], consisting of 1000 original video sequences that have been manipulated with four automated face manipulation methods: Deepfakes, Face2Face, FaceSwap, and NeuralTextures. The data have been sourced from 977 YouTube videos, and all videos contain a trackable mostly frontal face without occlusions which enables automated tampering methods to generate realistic forgeries. Owing to the Face2Face and NeuralTextures method in the FF++ dataset does not tamper the whole face (we obtain biological signals from the whole face, and when the tampering part is too small, it will reduce effectiveness of the method), we mainly verify our method on Deepfakes and FaceSwap datasets.

4.1.2. DeepFake Detection. The DFD dataset contains 363 original videos performed by actors and 3068 manipulated videos. These actors are required to perform different actions and then implement face-swap technology between different actors. To better extract the biological signals from the face, we chose a few specific actions, such as "podium speech happy" and "talking still." In these actions, the face is well facing the camera, and there are not too many interference factors. Therefore, we used 176 real videos and 754 fake videos from DFD. The biggest problem with the DFD dataset is the imbalance of positive and negative samples. So, we should expand the real video. The principle of expansion is any segment of the real video also is a real video. So we use the idea of sliding window to generate ppg_map. When processing real video, the stride of the sliding window is smaller than the length of sliding window, as shown in Figure 6. After expansion, we are equivalent to using 704 real videos and 754 fake videos.

4.1.3. UADFV. The UADFV dataset is proposed by Yang et al. [21], which contains 49 real videos and 49 fake videos. The average length of each video is about 11 seconds, and the resolution is 294×500 pixels.

4.2. Experiment Setting and Results. For generating ppg_map, we divide the face frame into $8 * 8$ blocks ($N = 64$) and used 64 frames ($T = 64$) to generate a ppg_map (which means the length of sliding window is 64), so the pixels of each ppg_maps are 64×64. Figure 7 shows a schematic of ppg_map.

We implemented this code on a workstation with four 2080Ti GPU cards. The model was trained using RMSprop for 160 epochs with a learning rate of 0.0004.

We used the Deepfakes dataset in FF++ (RGB dimension) to verify the effectiveness of the model. The accuracy and loss values of this model on the training set are shown in Figure 8. It can be seen from Figure 8 that as the epochs increase, the classification accuracy of the model is gradually increased, while the loss value is gradually decreased, and it stabilizes at 160 epochs, which illustrates the effectiveness of the model in this paper.

In order to prove the advantage of multidimensional signals, we analyzed the classification accuracy of single-dimensional signals and multidimensional signals, as shown in Table 1. The accuracy can be improved obviously when using multidimensional (M-D) signals.

4.3. Comparison. In order to verify the effectiveness of the method, a comparative experiment was carried out with the model mentioned in FaceForensics++, and the comparison results are shown in Table 2. The results show that our method has higher detection accuracy than other methods.

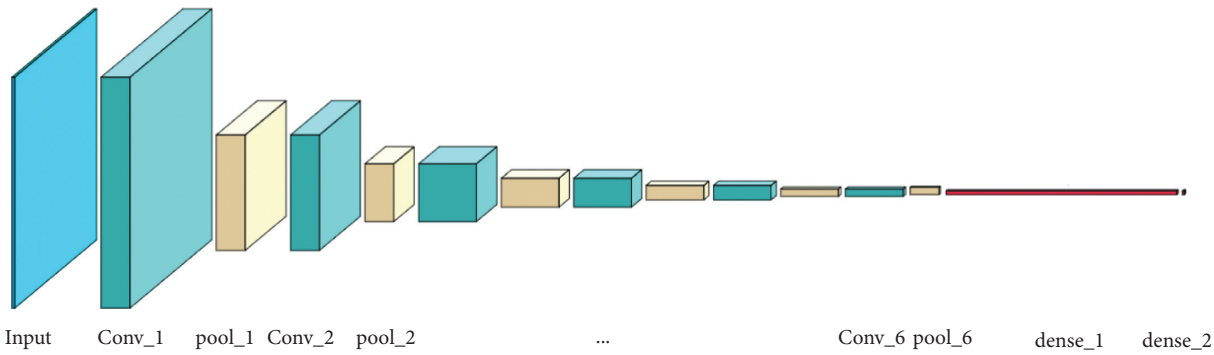

Figure 5: CNN architecture. We used six convolution layers with max pooling, followed by a flatten layer and dense layers.

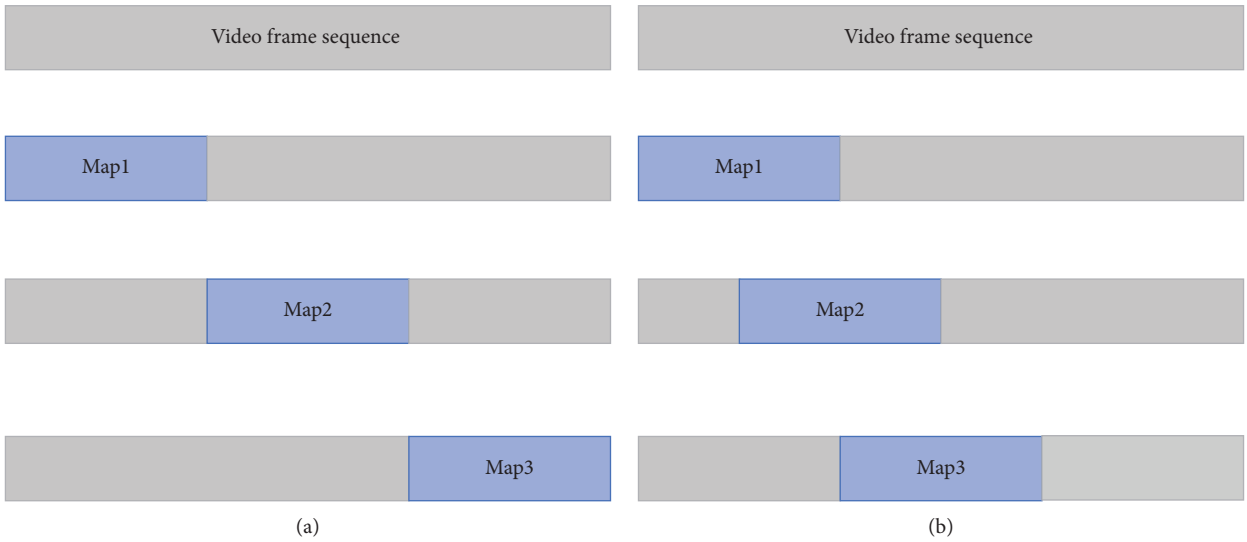

Figure 6: Different ways to generate ppg_map for real and fake videos. (a) When dealing with fake video, the stride of the sliding window is equal to the length of the sliding window. (b) When dealing with a real video, the stride of the sliding window is smaller than the length of the sliding window.

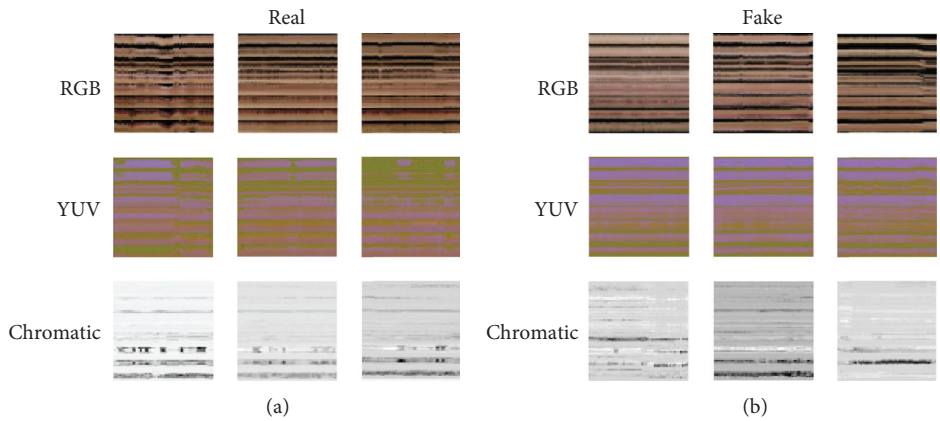

Figure 7: Schematic diagram of the ppg_map. (a) The ppg_maps generated by real videos. (b) The ppg_maps generated by fake videos.

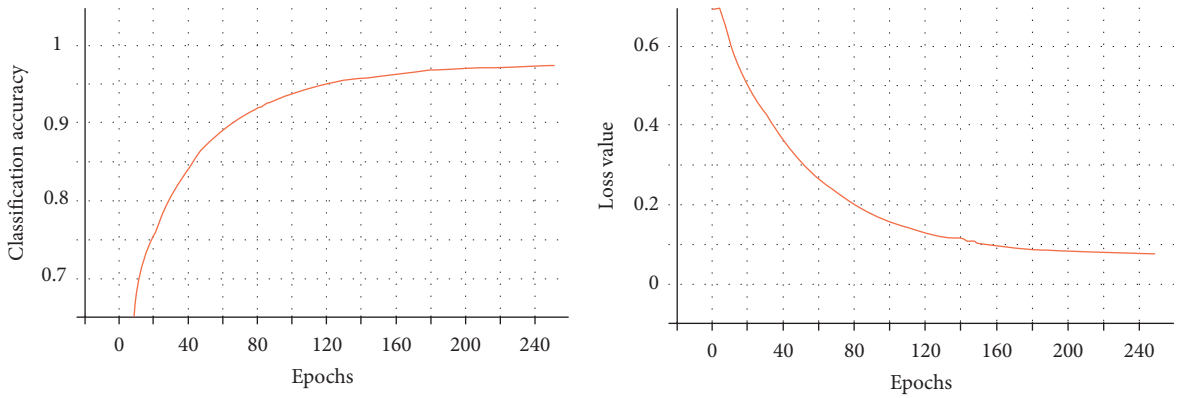

Figure 8: ((a), (b)) The changing curve of accuracy rates and loss value with training times, respectively.

Table 1: The accuracy on the testing set on different datasets.

Method	DFD (%)	UADFV (%)	Deepfakes (%)	FaceSwap (%)
Chrominance	86.74	87.23	81.23	79.25
YUV	90.21	89.36	94.23	88.26
BGR	94.26	88.30	96.20	90.84
M-D	**97.14**	**94.72**	**98.01**	**93.64**

Table 2: Comparison of experimental accuracy results by different models.

Method	Deepfakes (%)	Face2Face (%)
Afchar et al. [12]	87.3	56.2
Bayar and Stamm [27]	84.5	73.7
Rahmouni et al. [28]	85.5	64.2
Baek et al. [29]	71.8	68.6
Rossler et al. [30]	96.4	86.9
Dogonadze [31]	93.6	83.9
Ours	**98.01**	**93.94**

5. Conclusions

In this paper, we propose a forensics method based on biological signals, through a deep neural network to realize the classification of real and fake videos. The deepfake cannot effectively retain the biological signals in the face video. Consequently, we use multidimensional biological signals to analyze the differences between real and fake videos. However, some deepfake videos are hard to be exposed under the complicated conditions such as unstable character movements and complex scene switching. We hope that the deepfake detection in these scenarios could be solved effectively by using signal enhancement and denoising in the near future work.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China NSFC (grant numbers 62072343, U1736211), the National Key Research Development Program of China (grant numbers 2019QY(Y) 0206). The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements.

References

[1] P. Korshunov and S. Marcel, "Deepfakes: a New Threat to Face recognition? assessment and detection," 2018, https://arxiv.org/abs/812.08685.

[2] U. A. Ciftci, I. Demir, and L. Yin, "Fakecatcher: detection of synthetic portrait videos using biological signals," 2020, http://arxiv.org/abs/1901.02212.

[3] V. Conotter, E. Bodnari, G. Boato, and H. Farid, "Physiologically-based detection of computer generated faces in video," in *Proceedings of the 2014 IEEE International Conference on Image Processing (ICIP)*, pp. 248–252, IEEE, Paris, France, October 2014.

[4] P. V. Rouast, M. T. P. Adam, R. Chiong, D. Cornforth, and E. Lux, "Remote heart rate measurement using low-cost RGB face video: a technical literature review," *Frontiers of Computer Science*, vol. 12, no. 5, pp. 858–872, 2018.

[5] S. Tulyakov, X. Alameda-Pineda, E. Ricci, L. Yin, J. F. Cohn, and N. Sebe, "Self-adaptive matrix completion for heart rate estimation from face videos under realistic conditions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2396–2404, Las Vegas, NV, USA, June 2016.

[6] G. Balakrishnan, F. Durand, and J. Guttag, "Detecting pulse from head motions in video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3430–3437, Portland, OR, USA, June 2013.

[7] X. Niu, S. Shan, H. Han, and X. Chen, "Rhythmnet: end-to-end heart rate estimation from face via spatial-temporal

representation," *IEEE Transactions on Image Processing*, vol. 29, pp. 2409–2423, 2019.

[8] 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*, p. 1, 2020.

[9] B. Chen, X. Qi, Y. Zhou, G. Yang, Y. Zheng, and B. Xiao, "Image splicing localization using residual image and residual-based fully convolutional network," *Journal of Visual Communication and Image Representation*, vol. 73, Article ID 102967, 2020.

[10] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Capsule-forensics: using capsule networks to detect forged images and videos," in *Proceedings of the ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2307–2311, IEEE, May 2019.

[11] N. T. Do, I. S. Na, and S. H. Kim, "Forensics face detection from gans using convolutional neural network," 2018.

[12] 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, IEEE, Hong Kong, China, 2018 December.

[13] N. Bonettini, E. D. Cannas, S. Mandelli, L. Bondi, P. Bestagini, and S. Tubaro, "Video face manipulation detection through ensemble of CNNs," 2020, http://arxiv.org/abs/2004.07676.

[14] Y. Li and S. Lyu, "Exposing deepfake videos by detecting face warping artifacts," 2018, http://arxXiv.org/abs/:1811.00656.

[15] H. Zhao et al., "Multi-attentional deepfake detection," arXiv preprint arXiv:2103.02406 (2021).

[16] H. Liu, W. Zhou, D. Chen, T. Wei, W. Zhang, and N. Yu, "Spatial-phase shallow learning: rethinking face forgery detection in frequency domain," 2021, http://arxXiv.org/abs/:2103.01856.

[17] D. Güera and E. J. Delp, "Deepfake video detection using recurrent neural networks," in *Proceedings of the 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pp. 1–6, IEEE, Auckland, New Zealand, 2018, November.

[18] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan, "Recurrent convolutional strategies for face manipulation detection in videos," *Interfaces (GUI)*, vol. 3, no. 1, 2019.

[19] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, and H. Li, "Protecting world leaders against deep fakes," in *Proceedingd of the CVPR Workshops*, pp. 38–45, Venice, Italy, 2019, June.

[20] Y. Li, M. C. Chang, and S. Lyu, "Ictu oculi: exposing ai created fake videos by detecting eye blinking," in *Proceedings of the 2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pp. 1–7, IEEE, Hong Kong, China, 2018, December.

[21] 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, IEEE, Brighton, UK, 2019, May.

[22] L. Li, J. Bao, T. Zhang et al., "Face x-ray for more general face forgery detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5001–5010, Seattle, WA, USA, August 2020.

[23] R. Wang, F. Juefei-Xu, L. Ma, and X. Xie, "FakeSpotter: a simple yet robust baseline for spotting AI-synthesized fake faces," 2019, http://arxXiv.org/abs/:1909.06122.

[24] S. K. A. Prakash and C. S. Tucker, "Bounded Kalman filter method for motion-robust, non-contact heart rate estimation," *Biomedical Optics Express*, vol. 9, no. 2, pp. 873–897, 2018.

[25] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR*, IEEE, Kauai, HA, USA, 2001, December.

[26] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics: a large-scale video dataset for forgery detection in human faces," 2018, http://arxXiv.org/abs/:1803.09179.

[27] 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*, pp. 5–10, Web Tokyo, Japan, 2016, June.

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

[29] J.-Y. Baek, Y.-S. Yoo, and S.-H. Bae, "Generative adversarial ensemble learning for face forensics," *IEEE Access*, vol. 8, pp. 45421–45431, 2020.

[30] A. Rossler, D. Cozzollino, L. Verdoliva, C. Riess, J. Thies, and M. Niessnar, "Faceforensics++: learning to detect manipulated facial images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Seoul, Korea, February 2019.

[31] N. Dogonadze, O. Jana, and Ji Hou, "Deep face forgery detection," 2020, http://arxiv.org/abs/2004.11804.