2022 9th International Conference on Digital Home (ICDH) | 978-1-6654-5478-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICDH57206.2022.00046

DEEPFAKE DETECTION METHOD BASED ON FACE EDGE BANDS

Zhengjie Deng[*1,2],*Bao Zhang*[1,2] , *Shuqian He*[1],*Yizhen Wang*[3]

[1]School of Information Science and Technology, Hainan Normal University, Haikou,China.
[2]Guangxi Key Laboratory of Image and Graphic Intelligent Processing, Guilin, China.
[3]School of physics and Electronic Engineering, Hainan Normal University, Haikou,China.
email:hsdengzj@163.com,1321226352@qq.com,76005796@qq.com,wangxuesu1980@163.com

ABSTRACT

The rapid development of face forgery technology and the generation of realistic fake videos have caused serious harm to individuals, society and even the country, so it is important to detect deepfake videos. There are many detection methods available for forged videos, but the overall performance is yet to be improved and does not cope well with high quality forged images or videos. Observing that existing forgery algorithms leave synthetic forgery traces at the edges of faces when creating videos, this paper proposes a new method for detecting forged videos. It first finds the face edges from the video frames, then extracts the face edge bands as deep learning inputs and trains them based on EfficientNet-B3 to achieve effective detection of deepfake videos. Experiments show that the method in this paper can achieve more than 99.8% AUC values on all four forgery methods of the FaceForensics++ dataset.

Index Terms— deepfake, detection, face edge, EfficientNet

1. INTRODUCTION

In the past few years, artificial intelligence has developed rapidly, especially in the field of deep learning. Deep face forgery is becoming more and more sophisticated and has lowered the threshold for novice users, and with it, more and more forged images and videos are spreading on the Internet. Deep face forgery is simply the ability to tamper with a face in an image or video, replacing person A with person B.

More and more negative issues are emerging, such as fraudsters sending fake images and videos to victims and scamming them. Unscrupulous individuals have damaged the reputation of public figures by replacing the faces of well-known actresses with pornographic images or videos. There have also been malicious face swaps of national leaders and false news releases. This is no longer just a spoof against individuals, but seriously affects the security of society and the country.

In view of the pernicious effects of face forgery, we investigate the detection of forged faces, and the main work and contributions of this paper are as follows.

(1) For face forgery videos, this paper focuses on the edges of face synthesis for research, and proposes a new method for detecting forged faces, unlike previous methods that use complete face images for training.

(2) The method in this paper uses only pixels from the edge band of the face, using less data than other schemes, allowing the model to learn more specific feature information about the synthetic face edges, while also reducing the interference of extraneous elements from background information.

(3) This paper uses the more advanced EfficientNet as the training network, and experimental results show that this method achieves superior detection results on all four datasets of FaceForensics++.

2. RELATED WORK

Existing deep face forgery techniques consist mainly of Autoencoder and GAN. Autoencoder consists of an encoder and a decoder, where the face image is put into the encoder to learn the features of the face, and then the decoder restores the face; GAN consists of a generator and a discriminator, where the two play against each other, continuously reducing losses to improve themselves based on each other's feedback, and finally reaching a Nash equilibrium. The basic principle of generating deepfake is shown in Figure 1. The training process is two autoencoders, in which the encoders share weights among themselves, and the two decoders reconstruct face A and face B respectively, while the forgery process is to connect the feature set of face A to decoder B to achieve the reconstruction of face A to face B.

*:Corresponding author.

Acknowledgement. This work was financially supported by Science and Technology Project of Haikou (No. 2020-053, 2020-014, 2020-044), Hainan Natural Science Foundation (620RC604), the Open Funds from Guilin University of Electronic Technology, Guangxi Key Laboratory of Image and Graphic Intelligent Processing (GIIP2012), the National Natural Science Foundation (61502127), and Key R&D Projects in Hainan Province (ZDYF2019010).

978-1-6654-5478-0/22/$31.00 ©2022 IEEE
DOI 10.1109/ICDH57206.2022.00046

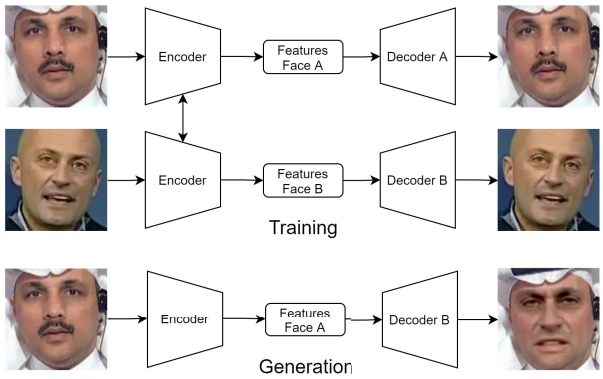

Fig. 1. The basic principle of Deepfake technology

There are currently open source forgery algorithms such as FAKEAPP, FaceSwap, DeepFaceLab and SimSwap to facilitate users to create their own forged videos, all of which rely on a large number of training images and better GPU resources.

In the face of the growing number of faked face videos, there has been widespread academic interest. More and more detection methods are being proposed, and current detection methods can be divided into three main categories of methods: manual feature-based, biometric feature-based and neural network feature extraction-based.

The first category of detection method focuses on the idea of image forensics, such as Li et al. [1] who proposed an adaptive frequency feature generation module to mine frequency information by separating the high and low frequency signals of video frames through discrete cosine transform and then recombining them before extracting the frequency features through convolution and linear pooling operations. Liu et al. [2] used the changes in the frequency domain during upsampling and proposed that the phase spectrum of real and faked videos that the phase spectrum in the frequency domain changes more significantly than the amplitude spectrum, allowing the model to focus on the learning of the phase spectrum. The first category of method only focuses on the frequency domain information of the image, but ignores the spatial information of the image.

The second category of detection method focuses on biological features, such as blinking, posture, mouth shape, pupil colour, artifacts, and other anomalies. Korshunov et al. [3] used both video and audio information to screen for tampering by comparing the differences in lip movements and sound matches between the real and fake videos; Agarwal et al. [4] pointed out that each person has a unique movement pattern, and face swapping leads to a mismatch between the movement patterns of the target and source objects. regions such as forehead, cheeks, nose, etc. to extract features for classification judgments. Chai et al [5] propose that local patches can be used to distinguish redundant artifacts of forged faces and that full convolution methods are used to train classifiers to focus on blocks of images, and this approach can be well generalised to different network architectures and image datasets. Li et al. [6] focus on the fact that forgery algorithms create boundaries in the face fusion step and propose Face-Xray method to predict the boundaries of face images and construct private datasets for experimental evaluation. The second category of detection method is suitable for early low-quality datasets, but as forgery techniques continue to evolve, the differences in these biometric features become less and less obvious, and the detection effectiveness decreases with them.

The third category of detection method focuses on learning face images by using existing or building their own deep neural networks to extract higher dimensional semantic features of faces for classification. SSTNet [7] combines spatial, steganalysis and temporal features for detecting DeepFakes. Specifically, the deep model XceptionNet [8] is used to extract spatial features, a simplified XceptionNet learns statistical features of image pixels under the constraints of traditional filters for steganalysis feature extraction, and RNN is used to extract temporal features. The FDFtNet approach [9] provides a fine-tuned network that can be used to improve the performance of existing CNN models in effectively detecting fake images and designed a fine-tuned transformer with self-attentiveness to extract different features from the images, and then MBblockV3 uses different convolutional and structure techniques to extract features. Rössler et al. [10] used the Xception network for pre-training can achieve good detection rates on separate datasets. Afchar et al. [11] built a Mesonet network to learn micro features of real and fake faces. nguyen et al. [12] built their own capsule network architecture to classify the features extracted by the VGG network. Khalil et al. [13] used HRNet networks for feature extraction and classification. The third category of method is a little better than the first two in terms of detection, but the generalisation of most of them needs to be improved.

3. DEEPFAKE VIDEO DETECTION METHOD

Deepfake is mainly tampered with the face, so we only need to analyze and experiment with the face area of the human face. This can not only focus on the more specific characteristics of the face area, but also reduce the interference factors from the background. The overall architecture of the detection model in this article includes the data pre-processing module and deep learning module. As shown in Figure 2, we first cut the video as a picture set, and then cut the face part according to the rectangular box obtained by the face detection tool. We extract the edge band of the face according to the method proposed, enhance the data of the face of the face, and then put it in the deep neural network for training.Finally, we use the softmax output image whether it is true and false.

As shown in Figure 3, this paper proposes a novel method

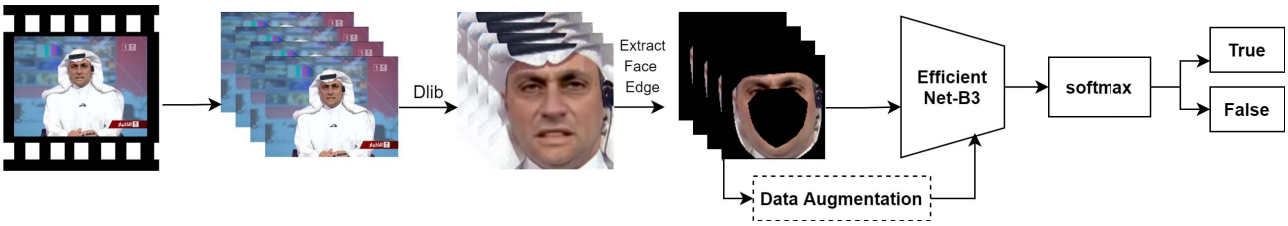

Fig. 2. The flow of the detection model proposed in this paper

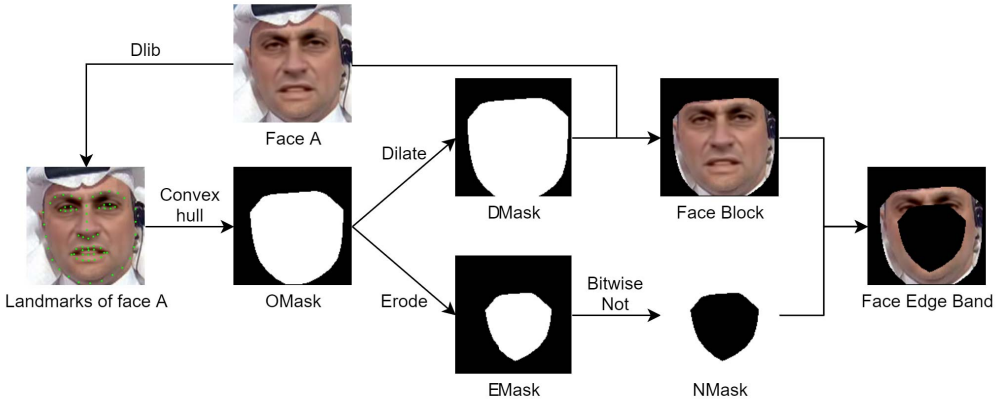

Fig. 3. The process of extracting face edge bands

to extract the face edge bands, and a detailed description of this paper's pre-processing method will be presented in Section 3.1.

3.1. Data pre-processing module

3.1.1. Face Extraction

In this paper, the dlib [14] toolkit was chosen for face detection and extraction. In the experiments, it was found that the face rectangle region positioned by dlib could not completely cover the face region, so the length and width of the face rectangle region were enlarged by 1.3 times respectively during cropping, and the final result is shown in Figure 4. The image was cut according to the enlarged face rectangle box for subsequent use.

Fig. 4. Face Extraction Flow Chart

3.1.2. Convex Hull Algorithm

Convex Hull is a minimal convex polygon that can contain all the points in the point set Q, where the points can be on the edges of the convex polygon or in the interior region [15]. The face in Figure 5 is labelled with the ordinal numbers of 68 feature points using dlib. In this paper, only 27 feature points from the two parts of the cheeks and eyebrows are needed to form the point set Q={p1,p2,...,p27}. In this paper, we use the convex hull algorithm on the face to obtain OMask.

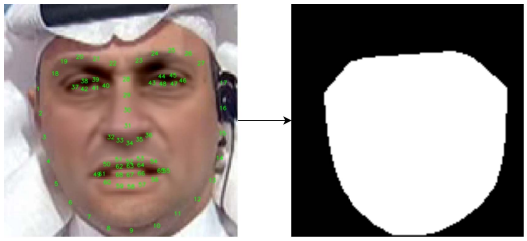

Fig. 5. Flowchart for generating OMask

3.1.3. Dilation Algorithm

In this paper, the dilation algorithm is used to scale the white area of OMask to obtain the DMask. The schematic diagram for mask dilation is shown in Figure 6, in which the sub-image

a is the original image and the sub-image b is the structuring element. By centering b with a, modifying the value of all pixels covered by b to 1, and b is overlaid with each pixel in a in turn, resulting in an inflated image c. In the experiment, this method sets the size of b not to be fixed, but to be scaled according to the size of the face image, set to 0.07 times the size of the face image.

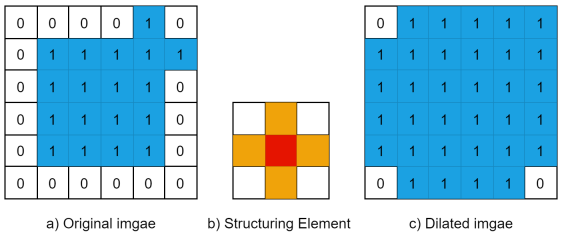

Fig. 6. Schematic diagram of the dilation algorithm

3.1.4. Erosion Algorithm

In this paper, the erosion algorithm is used to scale down the white area of OMask to obtain an eroded EMask. as shown in Figure 7, the reduction and expansion of the image is the opposite process, again using the structuring element b, overlapping b with the original image a, keeping the central pixel point in a that can contain all of b, and modifying the rest of the pixel values to 0. In this experiment, similar to the dilation algorithm, the size of b is scaled according to the size of the face image, set to 0.16 times the size of the face image.

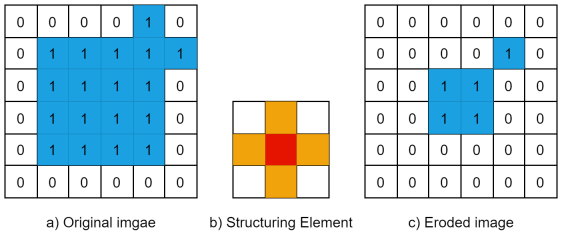

Fig. 7. Schematic diagram of the erosion algorithm

3.1.5. Bitwise Not Algorithm

The principle of inversion by bit is shown in Figure 8, where all the pixel values in a are changed from 0 to 1, and from 1 to 0 , to finally obtain b. In the experiment, the black and white area of EMask was inverted to obtain NMask.

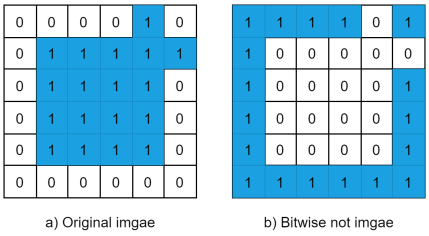

Fig. 8. Schematic diagram of the bitwise not algorithm

3.2. Deep learning module

The core idea of EfficientNet [16], which is referenced in this paper and released by Google in 2019, is to perform network search on width, depth, and resolution, and to perform compound scaling on the variables of the three dimensions, in order to pursue a balance between speed and accuracy while improving network performance. The base network structure of EfficientNet is EfficientNet-B0, and the compound scaling on this benchmark network yields the network structure of EfficientNetB1-B7. In order to balance training time and accuracy, the intermediate EfficientNet-B3 is chosen as the backbone network for training in this paper.

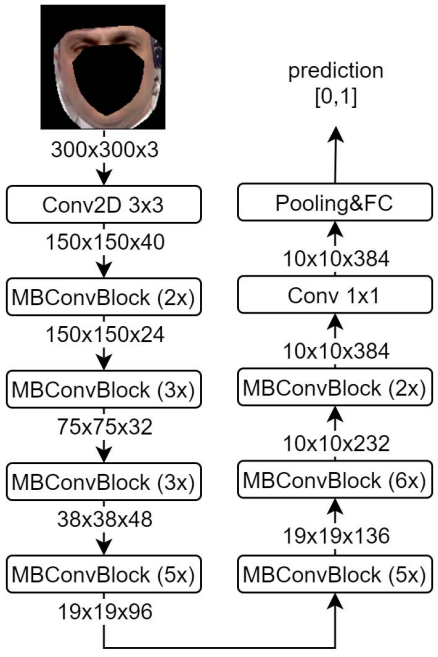

Fig. 9. The network architecture of EfficientNet-B3

The structure of EfficientNet-B3 is illustrated in Figure 9, which shows the size of the input and output, and the structure of EfficientNet-B3 is divided into three main parts. The first

part is a stem, which is used for initial feature extraction and consists mainly of a convolution, normalisation and activation function.The second part consists of 7 blocks with different numbers of sub-blocks, for a total of 26 sub-blocks. This part is the unique feature extraction structure of EfficientNet, which performs efficient feature extraction by stacking Blocks.The third part consists of a convolution, global average pooling and a fully connected layer.

4. EXPERIMENT

4.1. Datasets

In this paper, experiments were conducted using the most compared public dataset in the industry, FaceForensics++ [12], a large-scale forged face video dataset containing 1000 news videos collected from video websites. The dataset was tampered with using four forgery algorithms, Deepfake, FaceSwap, Face2Face, and NeuralTextures, to generate 1000 forged videos each. FaceSwap and Face2Face are computer graphics-based methods, while Deepfake and NeuralTextures are deep learning-based methods. All videos are available in three compressed versions, where Raw, C23 and C40 denote uncompressed, compressed parameter 23, and compressed parameter 40 respectively, and the picture quality is from high to low. Due to the large number of datasets, this paper chooses to conduct experiments on the C23 version.

Table 1. Distribution of the number of images in the dataset

Sub-dataset	Total	Train	Valid	Test
Original	170818	137628	16260	15924
DeepFake	168775	137573	16189	16010
FaceSwap	136399	109232	13137	13024
Face2Face	170369	137066	16434	15863
NeuralTextures	136159	108972	13151	13030

In this experiment, the dataset is divided into training set, validation set and test set in the ratio of 80%, 10% and 10%. In order to reduce duplicate pictures, take one in every three pictures, and Table 1 shows the exact number of images.

4.2. Experimental settings

This experimental platform is a Windows 10 operating system. The graphics card uses NVIDIA GeForce RTX 3090 with Intel (R) Core (TM) i7-10700K CPU@3.8GHz processor. The experimental code is implemented under the PyTorch 1.8 framework. The picture is adjusted to the best size 300x300 for the EfficientNet-B3 network before training. The training set is set up randomly flipping and normalization for data enhancement. In order to improve the classification effect, the migration learning strategy is used to initialize the model with the weight trained on ImageNet. This experiment uses a cross-entropy loss function and a learning rate adjustment strategy for the decline in random gradient. Setting batchsize is 32 and the number of iterations is 30 epochs

4.3. Experimental results

This experiment uses AUC to evaluate the effectiveness of the model. The ROC curve is used to demonstrate the performance of the classification model, using the false positive rate and the true positive rate as the horizontal and vertical axes respectively, reflecting the relationship between the two together.The AUC is the area under the ROC curve and is an important metric for evaluating the performance of the classifier, taking values between 0 and 1, with higher values indicating better performance.

Table 2. Compare the AUC of each detection algorithm on the test set of four forgery algorithms

Method	DeepFake	FaceSwap	Face2Face	NeuralTextures
Xception [10]	99.38%	99.36%	99.53%	99.50%
HRNet [13]	99.26%	99.24%	99.50%	98.61%
Face X-ray [6]	99.17%	99.20%	99.06%	98.93%
Ours	99.86%	99.91%	99.99%	99.87%

The results are shown in Table 2, which shows that the method achieves good detection results for all four forgery methods in the dataset. Compared with the other three benchmark methods, the AUC values of the methods in this paper are all improved, further demonstrating the effectiveness and superiority of the proposed methods in this paper.

5. CONCLUSION

The method proposed in this paper is very different from traditional pre-processing. Based on the synthetic edge defects of faked faces, this paper proposes a method to extract the face edge bands, and then put the face edge bands into EfficientNet-B3 for training and classification. This method uses only a small number of pixels in the face edge band, allowing the model to better learn the local feature information of the face edge band and reduce the information interference from the background. Compared with other methods, this method achieves an AUC of over 99.8% on all four datasets, outperforming the previous benchmark methods.

With the gradual development of artificial intelligence technology, new and better algorithms for faking faces will continue to emerge. In future work, we will continue to refine the method in this paper to further improve the performance of the detection.

6. REFERENCES

[1] Jiaming Li, Hongtao Xie, Jiahong Li, Zhongyuan Wang, and Yongdong Zhang, "Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 6458–6467.

[2] Honggu Liu, Xiaodan Li, Wenbo Zhou, Yuefeng Chen, Yuan He, Hui Xue, Weiming Zhang, and Nenghai Yu, "Spatial-phase shallow learning: rethinking face forgery detection in frequency domain," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 772–781.

[3] Pavel Korshunov and Sébastien Marcel, "Speaker inconsistency detection in tampered video," in *2018 26th European signal processing conference (EUSIPCO)*. IEEE, 2018, pp. 2375–2379.

[4] Shruti Agarwal, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li, "Protecting world leaders against deep fakes.," in *CVPR workshops*, 2019, vol. 1, p. 38.

[5] Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola, "What makes fake images detectable? understanding properties that generalize," in *European conference on computer vision*. Springer, 2020, pp. 103–120.

[6] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo, "Face x-ray for more general face forgery detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5001–5010.

[7] Xi Wu, Zhen Xie, YuTao Gao, and Yu Xiao, "Sstnet: Detecting manipulated faces through spatial, steganalysis and temporal features," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 2952–2956.

[8] François Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[9] Hyeonseong Jeon, Youngoh Bang, and Simon S Woo, "Fdftnet: Facing off fake images using fake detection fine-tuning network," in *IFIP international conference on ICT systems security and privacy protection*. Springer, 2020, pp. 416–430.

[10] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner, "Faceforensics++: Learning to detect manipulated facial images," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1–11.

[11] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen, "Mesonet: a compact facial video forgery detection network," in *2018 IEEE international workshop on information forensics and security (WIFS)*. IEEE, 2018, pp. 1–7.

[12] Huy H Nguyen, Junichi Yamagishi, and Isao Echizen, "Capsule-forensics: Using capsule networks to detect forged images and videos," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 2307–2311.

[13] Samar Samir Khalil, Sherin M Youssef, and Sherine Nagy Saleh, "icaps-dfake: An integrated capsule-based model for deepfake image and video detection," *Future Internet*, vol. 13, no. 4, pp. 93, 2021.

[14] Davis E King, "Dlib-ml: A machine learning toolkit," *The Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[15] David Avis, David Bremner, and Raimund Seidel, "How good are convex hull algorithms?," *Computational Geometry*, vol. 7, no. 5-6, pp. 265–301, 1997.

[16] Mingxing Tan and Quoc Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.