A Detection Method for DeepFake Hard Compressed Videos based on Super-resolution Reconstruction Using CNN

Zhang Hongmeng
Information Engineering University
Zhengzhou, China
+8617691113708
meng19950929@stu.xjtu.edu.cn

Zhu Zhiqiang
Information Engineering University
Zhengzhou, China
+8613607662989
xdzzqjs@163.com

Sun Lei
Information Engineering University
Zhengzhou, China
+8618537195378
sl0221@sina.com

Mao Xiuqing
Information Engineering University
Zhengzhou, China
+8613939096912
21166813@qq.com

Wang Yuehan
Information Engineering University
Zhengzhou, China
+8617334884899
Chineral@qq.com

ABSTRACT
The DeepFake video detection method based on convolutional neural networks has a poor performance in the dataset of hard compressed DeepFake video. And a large number of false tests will occur to the real data. To solve this problem, a networks model detection method for super-resolution reconstruction of DeepFake video is proposed. First of all, the face area of real data is processed by Gaussian blur, which is converted into negative data, and the real data and processing data are input into neural network for training. Then the residual network is used for super-resolution reconstruction of test data. Finally, the trained model is used to test the video after super-resolution reconstruction. Experiments show that the proposed method can reduce the false detection rate and improve the accuracy in detection of single frames.

CCS Concepts
• Computing methodologies→Artificial intelligence→ Computer vision→Computer vision problems→Object detection.

Keywords
Deep Learning; DeepFake detection; Super-resolution reconstruction; Hard compressed video

1. INTRODUCTION
In recent years, with the development of Internet of things, big data, cloud computing, and other new technologies, ubiquitous sensing data and graphics processor computing platform promote the rapid development of artificial intelligence technology represented by deep neural network. Driven by the three factors of algorithm, computing power and data, the third wave of development is coming [1]. In the past, the theory of artificial intelligence did not consider the open or even antagonistic system operating environment [2], which brought great opportunities, but also brought many risks. Because the application boundary of innovative technology is difficult to control, the abuse of artificial intelligence has gradually become prominent.

DeepFake is an important manifestation of the abuse of artificial intelligence. This concept first appeared at the end of 2017. A user of Reddit website named "deepfakes" published a pornographic video using FakeAPP to synthesize a star on the Internet, which attracted attention from all walks of life. DeepFake is a combination of "Deep Learning" and "Fake", which refers to the synthesis technology of human body image based on artificial intelligence, especially deep learning. With the development of technology, DeepFake technology has developed into multi-modal video deception technology including video forgery, voice forgery, text forgery and micro expression synthesis. The supporting technology behind DeepFake is Deep Learning, mainly including Generative Adversarial Networks (GAN) and Convolutional Neural Networks (CNN). The more the data of training deep learning algorithm, the higher the fidelity of synthesized audio and video products. In these fake videos and audio, people can say what they haven't said in reality, do what they haven't done in reality, and even reach the level of the fake and true, which impacts people's traditional cognition.

DeepFake is more and more harmful to data privacy and social security. The detection and filtering of forgery content also arises at the historic moment. Huawei, Alibaba, Google and other domestic and foreign artificial intelligence laboratories have put forward DeepFake video image detection schemes. Academic research on DeepFake video detection has gradually shifted from traditional detection to more robust and growing machine learning algorithm detection.

At present, DeepFake video has a variety of compression formats, and the detection methods are not robust and available, so it is difficult to be widely used in datasets with complex quality format, and the performance is worse in hard compression datasets. In this paper, a DeepFake detection method based on residual dual network model is proposed, which increases the availability of detection model and the robustness of detection algorithm. In view of the fact that the increase of artifact features after H.264 compression is the main factor affecting the accuracy of detection, this paper first uses residual network to reconstruct

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
HPCCT & BDAI 2020, July 3–6, 2020, Qingdao, China.
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-7560-3/20/07 …$15.00.
http://doi.org/10.1145/3409501.3409542

the super-resolution of the video to be tested, reduces the number of artifact features in the video frame, tests the reconstructed video input detection network, and improves the availability of the model in the hard compression datasets. Finally, the evaluation test is carried out on the standard dataset of FaceForensics++.

2. RELATED WORKS

Video detection method mainly through the analysis of the attributes, information, content and other data of video itself, find the rules to use for detection. This method can ensure that the original result of the video is not destroyed, and there is no special requirement for the source of the image and whether it has been processed. In the detection of DeepFake video, the research direction is mainly divided into two categories. The first category is based on the time characteristics across video frames to detect DeepFake. Most of the methods based on time characteristics use the depth learning recursive classification model. The second is the detection method of visual artifacts based on video frames, which is classified by deep or shallow depth neural network.

2.1 Detection Method based on Time Series

DeepFake algorithm couldn't effectively take time correlation into account in face changing. Sabir et al. use the temporal and spatial characteristics of video stream to detect forged video [3]. Video manipulation is carried out on the basis of frame by frame. Therefore, it can be considered that the low-level artifacts generated by manipulate at the top of the frame will further represent the time artifacts that are inconsistent across frames. Sabir et al. integrated the convolutional neural network DenseNet [4] and gated loop unit [5]. And proposed a cyclic convolution model (RCN) to capture the discontinuity between frames in the time domain. In the same way, Guera and Delp think that the face changing in DeepFake video contains the inconsistency of intra frame area and time between frames [6]. They proposed a time aware pipeline method, which uses CNN and long-term memory (LSTM) to detect DeepFake video. CNN is used to extract frame level features and feed them to LSTM to create time series descriptors. Finally, a full connection network is used to classify the real video and the fake video according to the sequence descriptor. On the other hand, Li et al. proposed to detect DeepFake video by blinking physiological signals [7]. They found that the blinking frequency of people in DeepFake video is much less than that of real video. But in the evolution of deep fake face changing video, this method is no longer applicable after consciously inserting blink datasets.

2.2 Detection Method based on Single Frame Features

When creating DeepFake face changing video, we need to use affine face transformation methods, such as scaling, rotating and clipping, to match the features of each region of the face in the source video. Due to the inconsistent resolution between the distorted face area and the surrounding environment, the process will produce artifacts that can be captured by CNN models, such as VGG16, ResNet50, ResNet101 and ResNet152 [8]. Zhou et al. proposed a dual flow framework to detect DeepFake Video [9], first using trained GoogleNet for face classification [10], the network stream learns the manipulated process generated artifacts, and classifies manipulation and real faces based on this feature. Then we train the other triple state flow, which captures the characteristics of the image patches with triple state loss, and uses the information clues hidden in the internal processing of the camera to detect the fake face. Finally, the final detection results are obtained by taking a weighted fusion score for the detection results of two network flows. However, the two network flows of this method are uncoordinated and uncorrelated, and the network flows of this method are not targeted to carry out structural design for relevant feature extraction. The final scheme evaluation is not based on the recognized datasets benchmark, which is not persuasive and robust. Afchar et al. proposed the MesoNet network [11] include meso-4 and mesoinception-4, has been tested on the standard dataset, but the feature extraction has not been fully explained. Only conjecture is put forward that the eye area is the difference feature in the real and fake images. Matern et al. trained a small fully connected neural network on the eyes, teeth, feature vectors of 16th dimension including eyes and teeth features and feature vectors extracted from the whole face, and fitted the logistic regression model as an example classifier to complete the classification of forged and real images [12]. However, this method is only applicable to the images that meet the preconditions such as opening eyes and exposing teeth. In addition, the test results depend on the specific test data, and the trained model does not have robustness. For example, the European and American human eye training model is not suitable for the test of fake wild face.

Li et al. proposed a deep learning method to detect DeepFake based on face detection artifact [13], which was put into CVPR2019 conference, and evaluated the proposed method on datasets UADFV and TIMIT. Li et al. compared the performance of this method with other methods, such as dual flow network, head-pose [14] and two DeepFake detection MesoNet networks. The advantage of this method is that it does not need to generate DeepFake video as negative samples before training the detection model. In order to improve the sensitivity of the feature, the method extracts the face region of the real image and aligns it, then applies Gaussian blur to the scaled image and generates the negative samples dynamically by the distorted affine transform. Compared with other methods, this method reduces a lot of time and computing resources, while other methods need to generate forged dataset for training or testing. This article provides a new idea for DeepFake video detection, but there is room for further the improvement in the video detection of the hard compression format. This is because after the hard compression, the data quality is reduced, and there are artifacts in the real dataset, resulting in CNN misjudging it. Therefore, this method first uses the residual network to reconstruct the super-resolution of the video to be tested to reduce the number of artifacts in the video frame, and then tests the reconstructed video input detection network to improve the availability of the model in the hard compression datasets.

3. METHOD

H. 264 video coding technology is the most widely used technology in video compression. Video in different compression formats in FaceForensics++ dataset is produced by this technology. In a frame image, for example, there is a great similarity and repeatability between the pixels in the face area, especially in the image with slow change, which is spatial redundancy. In the video sequence, there is a certain correlation between the content of adjacent frame images. In many cases, there is almost no difference in some pixel areas between adjacent frames, and even there is no change between some frames, which results in serious time redundancy. In the process of video coding, there are intra prediction, inter prediction and other steps, which is the mechanism introduced to solve the high redundancy of video signal [15]. Compression process is a process of reducing spatial and temporal redundancy, so it is also bound to introduce down sampling, fuzzy and quantization noise degradation process.

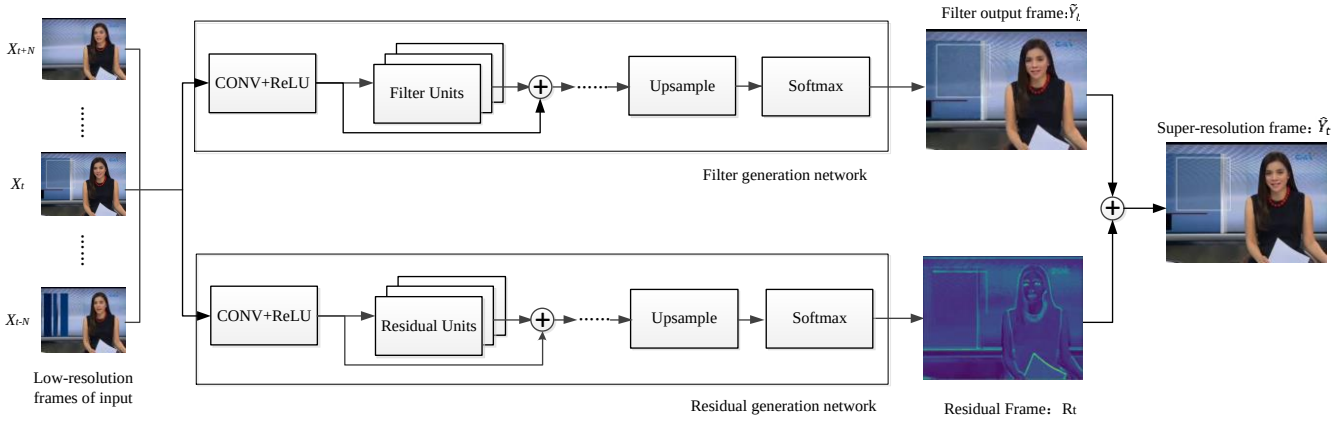

Figure 1. The network structure of video super-resolution.

Therefore, in the real dataset of hard compression format, the unique artifact features similar to the forged video frame will appear in the video, which will lead to the test results of the original detection model.

In order to solve these problems and improve the quality of compressed video restoration, super-resolution reconstruction of compressed video has gradually become a research hotspot. In the initial research, the traditional super-resolution reconstruction method is applied to video compression, but there are many disadvantages because of not considering the quality degradation of the compression process.

3.1 Video Super-Resolution Reconstruction

We use the method of Hyun et al. to use neural network to carry out super-resolution reconstruction of face video [16], which can reduce the artifact features in the real dataset and make the artifact features in the forged video not disappear. This reconstruction method implicitly uses the motion information to generate dynamic up sampling filter. Using the generated up-sampling filter, the HR frame can be directly constructed by local filtering of the input center frame, which can generate clearer and time consistent HR video.

Firstly, a filter generation network is used to construct the high definition frame directly by local filtering of the input central frame. The result of the network generated by the filter alone is lack of clarity, that is, the sharp details of the image. Because it is only the weighted sum of the input pixels, and some details couldn't be recovered through the linear filter. And because it does not rely on the explicit calculation of motion, nor directly combines the values in multiple frames. It can generate clearer and time consistent HR video. In order to solve this problem, high-frequency details are added through the prediction of residual generation network. The residual image is composed of multiple input frames rather than a single input frame, and the high-definition frame constructed by the network generated by the filter is used as the baseline, and then it is added with the calculated residual.

As shown in Fig.1, the network generates two outputs to generate the final HR frame $\hat{Y}_t$ based on a set of input LR frames $\{X_{t-N:t+N}\}$ dynamic up sampling filter F_t and residual R_t. Firstly, a dynamic up sampling filter F_t is used to locally filter the input central frame X_t and then the residual R_t is added to the up sampling result Y_t to get the final output $\hat{Y}_t$.

The filter and residual generation network share most of the weights, which can reduce the overhead of producing two different outputs. The parameter sharing of the network is inspired by the dense block, and has been modified to solve the problem of image super-resolution reconstruction. Specifically, 3D convolution layer is replaced by 2D to learn spatio-temporal features from video data. Because 3D convolution layer is more suitable for face motion recognition and temporal and spatial feature extraction of video data than 2D convolution layer. Each part of dense block is composed of batch normalization (BN), 1×1×1 convolution, BN, ReLU and 3×3×3 convolution. As described in the Fig.1, each section takes all previous feature maps as input. Each input LR frame is first processed by a shared 2D convolution layer and connected along the time axis. The generated spatiotemporal feature map is processed in a separate branch composed of 2D convolution layers after 3D dense blocks to generate two outputs.

3.2 DeepFake Detection

The detection algorithm uses the depth coding network to take the face region image as the input and output the corresponding image feature coding. The deep coding neural network mainly uses the

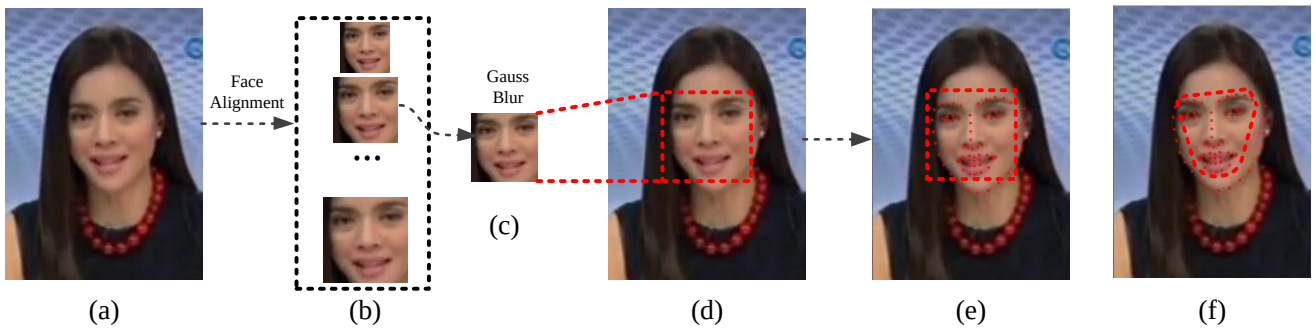

Figure 2. Negative data generation and RoI region enhancement.

convolutional neural network resnet50 model. The deep coding neural network obtains the feature coding directly as the input of the multi-layer fully connected neural network, thus forming the end-to-end neural network. Therefore, the BackPropagation algorithm can be used to train the network directly. In the training process, the parameters of the deep coding network are fine tuning with a smaller learning rate, and the parameters of the full connection layer are updated with a larger learning rate, so that it has a more suitable feature extraction ability for the face area, and finally achieves a higher classification accuracy.

In order to improve the ability of CNN to capture the artifact features and simplify the training process, the negative sample data is preprocessed by simulating the resolution inconsistency generated by the affine warpage transformation in DeepFake. First, we use dlib module to detect the face in the real data, then extract the face feature point coordinates to calculate the conversion matrix, align the extracted face to the same scale, and then use the Gaussian blur of 5 ×5 pixel kernel to smooth, so as to better simulate the discontinuity of each region of the forged data. In order to improve the diversity of training data, this scheme changes the image information of all training samples: brightness, contrast, distortion and sharpness. And as shown in the Fig.2 below, the region of interest (RoI) is created, that is, the rectangular region subtracts the convex polygon created according to the feature points at the bottom of the eyes and mouth.

Specifically, the coordinates of facial feature points are used to determine RoI, for example, $[y_0 - \widehat{y_0}, x_0 - \widehat{x_0}, y_1 - \widehat{y_1}, x_1 - \widehat{x_1}]$, where y_0, x_0, y_1, x_1 represents the minimum bounding box that can cover all facial feature point coordinates. The variable $\widehat{y_0}$, $\widehat{x_0}$, $\widehat{y_1}$, $\widehat{x_1}$ are random value between $\left[0, \frac{h}{5}\right]$ and $\left[0, \frac{w}{8}\right]$, where h, w are the height and width of the rectangular face respectively, and the RoI is adjusted to 224 × 224 to feed into CNN model for training.

During the training, negative examples and training process are generated in a dynamic way. For each training batch, half of the positive cases are randomly selected and transformed into negative cases according to the pre-processing method, which makes the training data more diverse. Use resnet50 model and fine tune it, set the batch training size to 64, the learning rate starts from 0.001 to 0.0001 in 1000 steps, and use SGD optimization method, the training process ends in the 20th iteration cycle, then use hard example mining strategy to fine tune the model, and finally use the trained parameter model to test the depth forgery data.

4. EXPERIMENTS

4.1 Experiments Setup

Rössler A and others expanded the face forgery detection dataset FaceForensics [17] and named it FaceForensics++ [18], and the dataset is based on DeepFakes, Face2Face, FaceSwap and Neural Textures, including a hidden test set and a database containing more than 1.8 million operable images. Google has released a large-scale visual DeepFake dataset [19] jointly produced by Google and jigsaw, which has been included in the FaceForensics++ dataset. As a standard dataset, the dataset has been widely used in the training and testing of DeepFake detection model.

In order to evaluate the performance of our improved detection method, this paper conducts experiments on DeepFake videos in the FaceForensics++ dataset. The DeepFake video library has about 1000 videos. Among them, the real face dataset used in the training set is the celeb dataset, and the fake face dataset is the DeepFake video database in the Face-Forensics++ dataset; the fake face dataset in the test set is the DeepFake video database in the FaceForensics++ dataset, and the real dataset is the DeepFake video database corresponding to the real face video database in the FaceForensics++ dataset. The experimental environment adopts the TensorFlow deep learning platform under the 64-bit windows 10 operating system.

4.2 Evaluations on Experiments

First of all, we verify the improvement of the detection accuracy of the real video frame in the hard compression dataset, randomly select the real and corresponding forged video in the hard compression format of FaceForensics++ to test, split the video into each frame to test, and output the probability that it is a forged video frame.

In this paper, the ROC curve is used to evaluate the proposed method. The full name of ROC is the "receiver operating characteristic" curve. Each point on the curve represents the response to the stimulus signal. Later, it was introduced into the field of machine learning to evaluate the classification and detection results. The samples are sorted according to the prediction results of the learner, and the samples are taken as positive examples one by one according to this order. Two important values (TPR and FPR) are calculated each time, and they are used as horizontal and vertical coordinates for drawing.

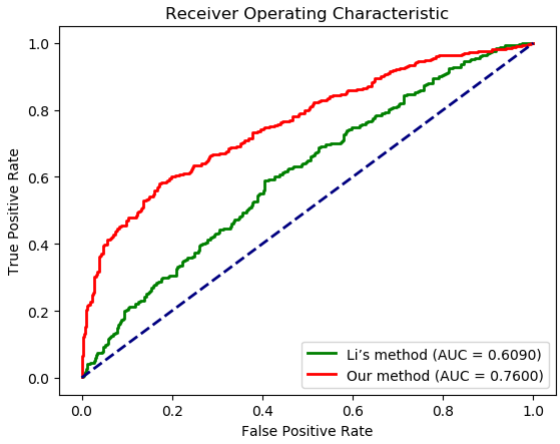

Figure 3. Performance of these method in the dataset of video frames.

Fig.3 shows the ROC curve comparison between the video frame of the classical method and the video frame after super-resolution reconstruction. In the experiment, the AUC value of Li et al. is 60.9%, while the AUC value of our method is 76%. Experiments show that our method has a better overall detection effect than Li et al.

At last, we verify the improvement of real video frame detection accuracy in the hard compression dataset. In the DeepFake video database of FaceForensics++ dataset, we randomly select 100 hard compression formats of real and corresponding forged video to test. In this paper, the number and degree of curve intersection of positive and negative samples are observed through the distribution curve of positive and negative samples to initially measure the classification ability of the detection model.

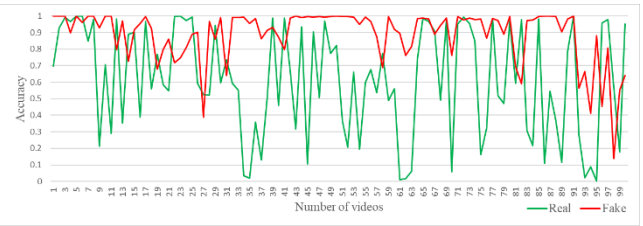

Figure 4. Performance of the original method in the dataset of hard compression format.

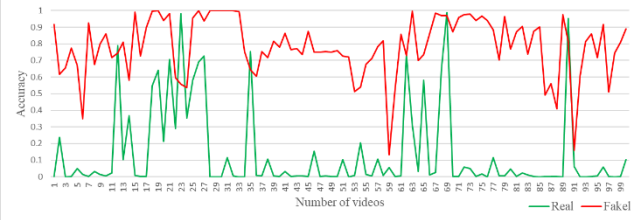

Figure 5. Performance of reconstruction method in dataset with hard compression format.

As shown in Fig.4 and Fig.5. Fig.4 shows the distribution curve of positive and negative sample detection accuracy of classical methods, and Fig.5 shows the distribution curve of our detection method based on super-resolution reconstruction. The horizontal axis of the graph represents the detected video number, and the vertical axis represents the probability that the video to be tested is judged to be forged. We can see that the curve of the real video in the classical method is relatively high, and there are many times of cross with the detection curve of the forged video, and the difference between the curves is small, so the preliminary evaluation of the detection model classification effect of the

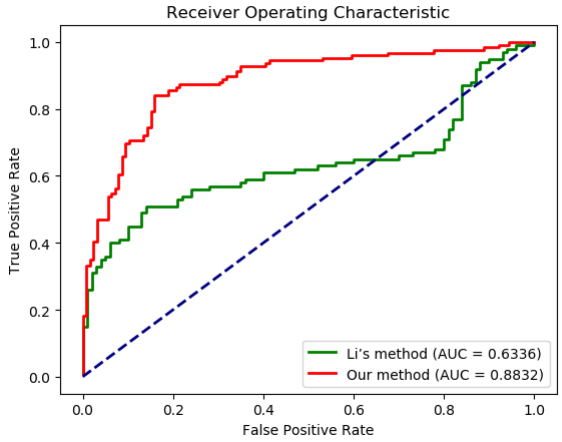

Figure 6. Performance of reconstruction method in dataset with hard compression format.

classical method couldn't meet our expectations. In our method in Fig.5, we can see that the curve of real video is relatively lower, and the cross times of positive and negative sample curves are significantly reduced, so it can be preliminarily determined that the detection effect of this method for hard compressed video is due to classic methods.

In this paper, based on AUC measurement, each frame of the real and corresponding forged video in 100 hard compression formats randomly selected from DeepFake video library in FaceForensics++ dataset is evaluated and reconstructed. The performance of the original method on the hard compression format video dataset is 63.36%, and the AUC value after reconstruction is 88.32%. The fig.6. results show that the method based on super-resolution reconstruction is effective to detect whether the hard compression video is forged, which can be used to improve the practicability of the original algorithm.

5. CONCLUSION

This paper presents a dual network model detection method for super-resolution reconstruction of DeepFake video. First of all, after aligning the face regions of real data, Gaussian blur is applied to process the face mosaic region, which is transformed into negative data to reduce the generation of negative data. Secondly, the real data and processing data are input into neural network for training, because of the preprocessing of previous data, the neural network is more sensitive to the extraction of face edge artifacts to improve the convergence speed of neural network effectively. Then use neural network to carry out super-resolution reconstruction of test data, neural network uses residual network to predict the difference between pixels of video frame in super-resolution reconstruction and increase the reconstruction details. Finally, use the trained depth network model to test the video after super-resolution reconstruction. The experiments results show that the proposed method reduces the error detection rate and has higher accuracy in the testing of single video frame. The disadvantage is that the detection performance of lossless compressed data is not good enough, which is also the key work to be studied in the future.

6. REFERENCES

[1] Wang, Fei-Yue, Ruqian Lu, and Daniel Zeng. "Artificial intelligence in China." IEEE Intelligent Systems, 2008, 23(6): 24-25.

[2] Li X, Zhang T. "An exploration on artificial intelligence application: From security, privacy and ethic perspective." 2017 IEEE 2nd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA). IEEE, 2017. pp. 416-420

[3] Sabir, Ekraam, et al. "Recurrent convolutional strategies for face manipulation detection in videos." Interfaces (GUI) 2019, 3: 1.

[4] Huang, Gao, et al. "Densely connected convolutional networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017, pp. 4700-4708.

[5] Cho K, van Merri ënboer B, Gulcehre C, et al. "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation." Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014, pp. 1724-1734.

[6] G üera D, Delp E J. "DeepFake video detection using recurrent neural networks". 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). IEEE, 2018, pp. 1-6.

[7] Li Y, Chang M C, "Lyu S. In ictu oculi: Exposing ai created fake videos by detecting eye blinking". 2018 IEEE International Workshop on Information Forensics and Security (WIFS). IEEE, 2018: 1-7.

[8] He, K., Zhang, X., Ren, S., Sun, J. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[9] Zhou, P., Han, X., Morariu, V., Davis, L. "Two-stream neural networks for tampered face detection". IEEE Computer Vision and Pattern Recognition Workshops. 2017, pp. 1831– 1839

[10] Szegedy C, Vanhoucke V, Ioffe S, et al. "Rethinking the inception architecture for computer vision ". Proceedings of the IEEE conference on computer vision and pattern recognition. 2016, pp. 2818-2826.

[11] Afchar D, Nozick V, Yamagishi J, et al. "Mesonet: a compact facial video forgery detection network". 2018 IEEE International Workshop on Information Forensics and Security (WIFS). IEEE, 2018, pp. 1-7.

[12] Matern F, Riess C, Stamminger M. "Exploiting visual artifacts to expose DeepFakes and face manipulations". 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW). IEEE, 2019, pp. 83-92.

[13] Li Y, Lyu S. "Exposing DeepFake Videos By Detecting Face Warping Artifacts ". Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2019, pp. 46-52.

[14] Yang X, Li Y, Lyu S. "Exposing deep fakes using inconsistent head poses". ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 8261-8265.

[15] Heiko Schwarz, Detlev Marpe, Thomas Wiegand. "Overview of the Scalable Video Coding Extension of the H.264/AVC Standard". IEEE Transactions on Circuits and Systems for Video Technology, 2007, 17(9) , pp. 1103-1120.

[16] Jo Y, Wug Oh S, Kang J, et al. "Deep video super-resolution network using dynamic upsampling filters without explicit motion compensation". Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018, pp. 3224-3232.

[17] Rössler A, Cozzolino D, Verdoliva L, et al. "FaceForensics: A large-scale video dataset for forgery detection in human faces". arXiv preprint arXiv:1803.09179, 2018.

[18] Rossler A, Cozzolino D, Verdoliva L, et al. "Faceforensics++: Learning to detect manipulated facial images". Proceedings of the IEEE International Conference on Computer Vision. 2019, pp. 1-11.

[19] Nick Dufour, Google Research and Andrew Gully, Jigsaw, Contributing Data to DeepFake Detection Research [EB/OL]. https://ai.googleblog.com/2019/09/contributing-data-to-DeepFake-detection.html.