Neural Processing Letters (2021) 53:4159–4175
https://doi.org/10.1007/s11063-021-10588-6

Exposing DeepFake Videos Using Attention Based Convolutional LSTM Network

Yishan Su[1] · Huawei Xia[1] · Qi Liang[2] · Weizhi Nie[1]

Accepted: 9 July 2021 / Published online: 4 September 2021

Abstract
The detection of face tampering in videos created by artificial intelligence techniques (commonly known as the *Deep Fakes*) has become an important and challenging task in network security defense. In this paper, we propose a novel attention-based deep fake video detection method, which captures the sharp changes in terms of the facial features caused by the composite video. We utilize the convolutional long short-term memory to extract both spatial and temporal information of DeeFake videos. Meanwhile, we apply the attention mechanism to emphasize the specific facial area of each video frame. Finally, we design a decoder to further fusion multiple frames information for more accurate detection results. Experimental results and comparisons with state-of-the-art methods demonstrate that our framework achieves superior performance.

Keywords DeepFake detection · Convolutional LSTM · Attention

1 Introduction

These days, digital videos have become much easier to obtain than before. With the rapid development of mobile photography equipment such as smartphones, people can easily make photos or videos containing facial data [45]. Social media like YouTube and Instagram enables us to share these data online, of which the content is far more abundant than traditional text messages [43]. The massive number of facial data [44] has evoked different kinds of facial synthesizing technology. Such a technique has the risk of being abused for malicious purposes

✉ Qi Liang
tjuliangqi@tju.edu.cn

Yishan Su
yishan.su@tju.edu.cn

Huawei Xia
xiahuawei@tju.edu.cn

Weizhi Nie
weizhinie@tju.edu.cn

[1] The School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China

[2] The School of Microelectronics, Tianjin University, Tianjin 300072, China

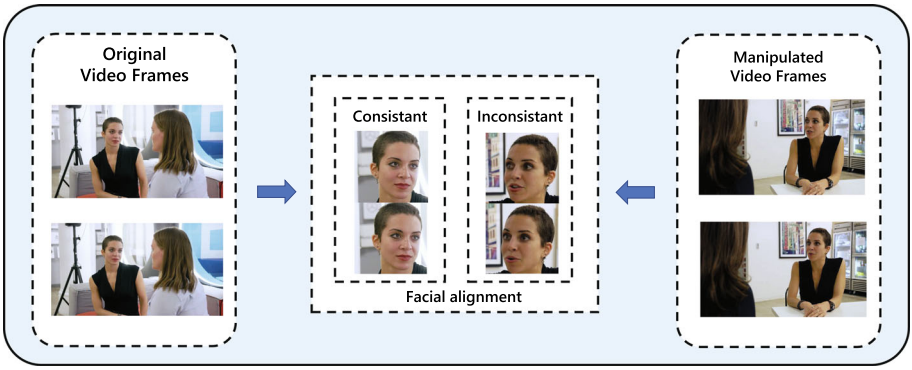

Fig. 1 The inconsistent facial feature across the manipulated video frames, note that the images are taken frame by frame from the target videos. We discovered that the facial feature changes in the manipulated video frames are much more pronounced than in the original video frames

like identity fraud [2,26], public opinion controlling [25], and so on. All these threats make the development of the corresponding detection method for the manipulated facial data necessary.

Among those facial synthesizing methods, a deep learning-based facial manipulation technique commonly known as DeepFake [34] has become very popular recently. Traditional DeepFake technique uses two autoencoders for parallel training, of which the outputs are often combined with notable features like the resolution changes in the manipulating area [15,48], the inconsistent of the head pose movements [11,41], and so on. These features are easy to be detected directly through our human eyes at first. However, with Generative Adversarial Networks [46] and related GAN-based optimizing methods (e.g. CycleGAN [50]) are put forward, the DeepFake technique can generate high-quality facial videos and images, these outputs are no longer be easily distinguished by people anymore, which makes the task of deep fake detection more challenging.

Among the existing DeepFake detecting methods, the MesoNet [1] focuses on the mesoscopic properties of images, whereas some of the features in DeepFakes (e.g. face warping artifacts) are not appropriately considered. Li et al. [16] proposed a CNN-based detecting method that focused on detecting face warping artifacts. This detection method mainly deals with still images, whereas the inter-frame temporal order information is often ignored. The RNN-based network [24] was then put forward to enforce the temporal information in DeepFake videos. Although the RNN-based network considers the temporal feature across frames, there are still drawbacks of discarding the facial feature incoherence among the input DeepFake video streams. We find that the sudden changes in facial expressions between the videos, or stiff changes, are obvious in comparison with the original videos. These characteristics are shown in Fig. 1. We can observe from the figure that the eyes, nose, and mouth of the human face in the manipulated video drastically change from frame to frame, but this change is not apparent in the original video. Thus, the inconsistent facial feature changes in DeepFake videos are concerned in our detection methods.

To alleviate the situation, we put forward a novel DeepFake videos detection method based on the weights of the input. The general processing structure of the network can be described as a ConvLSTM [27] network, which is based on an Encoder–Decoder structure that enforces the facial feature incoherence of Deepfake videos. In comparison with other methods, our design enables the network to focus on both the temporal and the spatial information of the input. The attention mechanism is then applied to remove redundant features and guide the

information fusion in the process of inter-frame feature aggregation between the encoder and the decoder. We conducted further experiments on the public data set FaceForensics++ [23] to prove the effectiveness of our proposed method.

The main contributions of this work can be summarized as follows:

- We propose a novel Encoder–Decoder structure to reinforce the feature across frames. Such a design enables the network to focus on the inconsistent changes of the facial feature across frames. The ConvLSTM model is applied to specialize both the temporal and the spatial information of the input.
- We combine the attention mechanism with the ConvLSTM model to enhance the region of interest in the video frames. Such a design enables our network to concentrate more on the facial features of the input.
- We verify our design through a series of experiments. The results show that our method has achieved state-of-the-art performance comparing with recent detecting methods based on the public dataset.

The rest of the paper is organized as follows: In Sect. 2, we concisely review the current related work in this field. Then, we describe the processing architecture of our network in Sect. 3. The corresponding experiments, results, and analysis are introduced in Sect. 4. Finally, in Sect. 5, we summarise and make our conclusion of this paper.

2 Related Work

Over the past decade, facial manipulation techniques are commonly used in fake video and image synthesizing [22]. The impact of fake videos and images became severe after [31] introduced a method that could generate real-time facial manipulation videos. Mobile applications like ZAO and FaceAPP enables attackers to generate a fake vision of the victim much easier than before. So it is necessary to develop facial manipulation detecting methods. In this section, we generally describe the facial manipulation technique and the detection of manual synthesized videos and images.

2.1 AI-Based Facial Manipulation Algorithms

Such technique can be divided into four categories, namely (1) entire face synthesis, (2) facial attributes manipulation, (3) face swap, (4) facial expression manipulation. The entire face synthesis technique [36] mainly focuses on creating a nonexistent face. With the successive appearance of deep learning methods [14], the entire face synthesis technique has the capacity to generate high-quality facial images which can hardly be distinguished through human eyes.

Facial attributes manipulation [38] mainly focuses on changing semantic aspects of the human face, for example, “hair”, “gender”, “skin color”, etc. Such a manipulation process based on the Convolutional Neural Network(CNN) [39] requires plenty of images for the training process and can only generate faces with different poses. With the introduction of GAN-based models e.g. Self-Perception GAN, the manipulation process becomes efficient and compact. Face swap can manipulate a person’s face by directly change it into facial images from someone else [48]. Traditional methods based on computer graphic calculation [34] and the deep learning-based methods known as DeepFakes are widely introduced in face-swapping manipulation. Applications like ZAO and FaceSwap can generate very authentic face-swapping videos. The facial expression can be modified through deep learning-based

models like Face2Face, which can modify the expression of the target person's face, e.g., changing a delightful expression into a depressed one.

2.2 Facial Image Synthesis Detecting Algorithms

Traditional methods mainly focus on particular facial manipulation evidence, which may have some limitations, for example, averaging the weights of the output channel to observe the difference between the manipulated images and the original ones [19]. Recently, deep learning-based facial image synthesis detecting models have made significant progress.

In order to avoid the probability of focusing on specific facial manipulation features and achieve robust manipulation detecting results [2], Zhou et al. [49] proposed a CNN-based two-stream neural network to capture tampering artifacts and local noise residual evidence of the image. Nguyen et al. [21] introduced a convolutional neural network using a semi-supervised learning method to detect and locate the manipulated regions in the target images. A Y-type decoder is introduced to conduct manipulation area segmentation and input reconstruction simultaneously.

2.3 Detection of Facial Manipulation Videos

The detection of facial manipulation videos is generally treated as a binary classification problem, which mainly consists of two aspects, one focuses on detecting the features across frames using recursive classification methods, the other detects the visual artifacts within frames, generally by introducing deep or shallow classifiers after the feature extraction.

In order to detect the temporal features across frames, Guera et al. [8] proposed a temporal-based system based on the CNN and LSTM networks. Li et al. [15] discovered that DeepFake videos rarely have the data within a person's eyes closed, thus can be detected through the blinking frequency of a video. The detecting networks based on CNN/RNN are commonly applied to capture the lack of eye blinking in these synthesized videos. However, such detecting methods can be avoided by deliberately inserting images with closing eyes during training. With limited computing resources and production time, the DeepFake algorithm can only generate videos and images of which the resolution has a severe decrease in comparison with the original input. In order to better fit the original inputs, these processed facial data must undergo a specific affine transformation. Li et al. [16] found out that this transformation process was generally combined with unique features in the generated Deepfakes videos. These artifacts can be detected by classical deep learning-based networks e.g. VGG, ResNet. The process of negative sample generation can be simplified by directly simulating the affine surface warping step. In fact, micro-feature detection methods based on noise analysis are hard to fit the environments where these features are significantly reduced.

3 Methods

The manipulated videos are detected by exploiting the facial feature incoherence resulted from the DeeFake video production pipeline. In this section, we will introduce our approach in detail. We start by introducing the network input, which is a sequence of preprocessed DeepFake video frames. These input frames are aligned and sampled in equal intervals to provide the facial data representing each video. Then we randomly enhance the input data by applying methods such as eye feature removal, facial area blurring, and so on. Such design

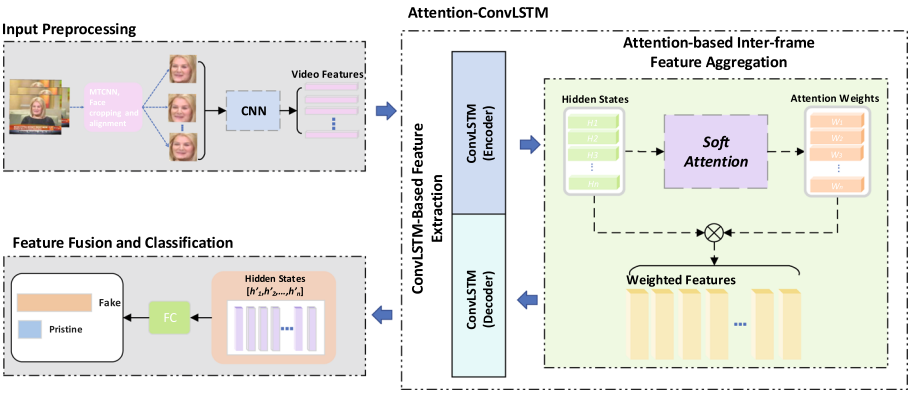

Fig. 2 The framework of our method. First, we detect and extract the faces in the video frames using the MTCNN model. Then we obtain the video features through a CNN network. The convolutional LSTM is then applied to generate the attention-based features. These features are then sent into another convolutional LSTM network, followed by a fully connected layer to generate the final classification results

enables the network to maximize the detailed facial features of the input. As a result, an input sequence of facial frames is generated from the models. The input frames are first passed through a CNN-based network to extract the visual features. These visual features are then fed into a ConvLSTM-based encoder to generate the weights of the sequential input data. We then apply the attention mechanism, which utilizes the weights to focus on the features that are more important for the task of making full use of the information carried by the input sequence. The output features are then sent into another ConvLSTM-based model as the decoder of the whole network. Such a design aims to detect the significant changes in facial features crossing the input data. The outputs are then used for the final classification task, where these features are fused to generate the classification results. We provide a detailed introduction in the following sections: (1) data preprocessing, (2) inter-frame feature aggregation, (3) feature fusion and classification.

3.1 Input Preprocessing

3.1.1 Facial Detection and Alignment

In order to extract the detailed facial data from each video, a deep cascaded multi-task framework MTCNN [47] is used to attain the landmarks of the target faces from the video frames. However, the facial landmarks gained from the MTCNN model have the probability of not covering the entire facial area from the original frames. Therefore, to solve this problem, we then take a dilation at a scale of thirty percent of the detected area. These processed video frames are then reshaped into the same size of (224×224). We randomly take 15 frames that are already aligned into the same scale of each video as the final input of our network. Therefore, an input sequence of 15 frames containing facial data which are aligned into the same size is generated to represent the target video.

3.1.2 Input Data Preprocessing

For the DeepFake video detection network, the larger the variety of the input videos is, the more information the input contains. As a result, the variation of the input is also an important parameter. We take the Gaussian Blur, Wavelet Transform, and facial data removal as the variation methods. The faces in the original videos are detected and extracted from the video frames by using the deep learning-based method MTCNN [47], followed by the Gaussian Blur with kernel size (5×5). The procedure aims to create more cases of resolution on distorted faces, which can better simulate inconsistent resolution changes in DeepFake videos. For comparison, we use the wavelet transform method to reduce image noise, thus improving classification performance under low-quality images. In order to imitate artifacts produced by the DeepFake generating pipeline, the modified faces generated by both methods are then affined and distorted to the original size of the input faces. To solve the problem of shrinking other essential features in DeeFake video generation e.g. the eye blinking and the details around the lips pose, we take frames from which the mouth, nose, and eyes are removed randomly while the rest remain unchanged. The distribution ratios of the manipulated frames to the original frames are all set to 1:2.

3.1.3 CNN-based Frame Feature Extraction

Since EfficientNet-B5 [30] has achieved a relatively better balance between the performance and the memory cost of the network compared with other traditional CNN models (e.g., AlexNet [13], VGG-Net [28]), the EfficientNet-B5 model is employed in our approach to extract the visual feature vector from the input frames. With the high efficiency of keeping the balance between the depth, width, and resolution of the network, EfficientNet can significantly improve and shorten both the accuracy and the training process in comparison with other traditional networks e.g. ResNet [9]. We take the scaled version of the baseline network EfficientNet-B0 [30] developed by the neural architecture search method [51]. Thus, the feature vectors $\{v_1, \ldots, v_n\}$ at the dimension of 2048 are used as the output of the model.

3.2 Attention-ConvLSTM

We propose a novel DeeFake video detection method based on the ConvLSTM model and the attention mechanism. Such a design is for extracting the temporal and facial information from the input frames while concentrating on the area of interest in the video frames. Our network architecture is shown in Fig. 2, where the details are discussed in the following subsections.

3.2.1 Convolutional LSTM Based Feature Extraction Network

We propose a purely neural network-based method to assign weights for the sequential structure of the input frames. Being widely used in many fields [7,17,42], the LSTM network [10] could take advantage of the features crossing the sequential data, which makes LSTM an effective solution for detecting the facial feature incoherence across video frames in the task of DeepFake video detection. The LSTM structure we apply is different from the classic ones applied in [7], it is a convolutional LSTM network based on an encoding-decoding structure. The architecture of the Convolutional LSTM-based feature extraction model is shown in Fig. 2. Compared to generic LSTM-based models [3,10,12], our network could optimize

the input by introducing weights generated by the ConvLSTM model. Such design enables the network to focus on features that are more important in the task of detecting DeeFake videos. Moreover, the ConvLSTM model could take advantage of both temporal and spatial information of the sequential input data, which is very efficient in processing video streams. We utilize the hidden state h of the encoding ConvLSTM structure to generate the weight w for the attention mechanism, h and w are then manipulated of which the output is sent into the decoding ConvLSTM structure to generate the final output h'. Such a design enables the network to concentrate on both the inter-frame and the intra-frame features in the task of DeepFake video detection.

3.2.2 Attention Based Inter-frame Feature Aggregation

Since we have obtained the feature vector v, the soft-attention mechanism [40] is applied in combination with the ConvLSTM model. Such a design is to specify the importance of the vector in which the feature is more significant in the encoder–decoder structure of our feature extraction network. The weights calculated by the attention mechanism are then multiplied with the hidden states H of which the results are then sent into the decoding network. The attention-based video weight is calculated as follows:

We apply the ConvLSTM network to exploit the structure information of the input weighted feature vectors $\{v_1, \ldots, v_n\}$. The general processing structure of the convolutional network we employ is inspired by Shi et al. [27]. Evolved from the RNN model, the convolutional model retains the hidden state h_t in the RNN and adds the cell state c_t to avoid information loss. The relationship between h_t and c_t can be expressed by the following formula:

$$h_t = o_t \bigodot tanh\left(c_t\right), \tag{1}$$

where $\bigodot$ denotes the Hadamard product. The output gate o_t is calculated as follows:

$$o_t = \sigma\left(W_{vo} * h_{t-1} + W_{ho} * v_{i,t} + b_o\right), \tag{2}$$

where σ represents the logic sigmoid function and $v_{i,t}$ stands for the feature vector corresponding to the tth selected frame in the ith video. W_v, W_h stands for the convolutional kernels, b represents the bais. The symbol $*$ stands for the convolution operator. The current memory state c_t results from the following formulation:

$$c_t = f_t \bigodot c_{t-1} + i_t \bigodot \widetilde{c_t}, \tag{3}$$

where c_{t-1} stands for the previous memory state and $\widetilde{c_t}$ represents the updated memory. $\widetilde{c_t}$ together with the input gates i_t and the forget gates f_t are calculated by:

$$f_t = \sigma\left(W_{vf} * h_{t-1} + W_{hf} * v_{i,t} + b_f\right) \tag{4}$$

$$i_t = \sigma\left(W_{vi} * h_{t-1} + W_{hi} * v_{i,t} + b_i\right) \tag{5}$$

$$\widetilde{c_t} = tanh(W_c * v_{i,t} + W_c * h_{t-1} + b_c). \tag{6}$$

With the advantage of enabling the neural network to focus on a series of particular portions of the input, the attention mechanism could significantly reduce the calculation cost and improve the network performance. We employ the soft attention mechanism in our work to minimize the task complexity, the processing structure can be described by the following formula:

$$e_i = W_m tanh\left(W_{va} * v_{i,t} + W_{ha} * h_{t-1} + b_a\right) \tag{7}$$

$$\alpha_i = exp\left(e_i\right) / \sum_{j=1}^{n} exp\left(e_j\right), s.t. \sum_{i=1}^{n} \alpha_i = 1 \tag{8}$$

$$\widetilde{v_t} = \alpha_i \cdot v_t, \tag{9}$$

where α_i represents the attention weight W_m, W_{va}, W_{ha}, and b_a are the parameters which are estimated together with the whole network. $\widetilde{v_t}$ stands for the updated input, which is then sent into the decoding ConvLSTM network, and get a set $H' = \left\{h_1{}', h_2{}', \ldots, h_n{}'\right\}$, which is used to represent the input frame sequence.

3.3 Feature Fusion and Classification

Since the detection of DeepFakes is a binary classification task, and the outputs of the ConvLSTM-based decoding model $h_n{}'$ are frame-level features, it is necessary to fuse these features for the task of classification. As the output features $h_n{}'$ generated by our encoding-decoding model have already been weighted by the attention mechanism, it is unnecessary to apply weights for feature fusion tasks. Therefore, we applied max pooling for the task of feature fusion. Such a design aims to fuse the weighted features with high efficiency while comprehensively represent the output of the encoding-decoding model. For classification, we applied a fully connected layer followed by a softmax layer to compute the probabilities of the input frame sequence being either pristine or Deepfake. This is the final descriptor of the entire video for the classification task.

During training, we adopt the cross-entropy loss to measure the accuracy of the prediction, the loss is calculated by the following formula:

$$L = -\sum_{\{V,c\} \in D} \left(c \cdot \log\left(\hat{c}\right) + (1 - c) \cdot \log\left(1 - \hat{c}\right)\right), \tag{10}$$

where $D = \{V, c\}$ represents the training dataset, V denotes the input videos, and c is a binary scalar specifying whether the video V is pristine or DeepFake. $\hat{c}$ is the predicted probability.

4 Experiments

In this section, we start by introducing the overall experiment setups and then present extensive experimental results to demonstrate the superiority of our method.

4.1 Dataset

The recently released benchmark dataset FaceForensics++ [23] (FF++) consisted of 1000 original videos was adopted in our experiments to evaluate the performance of our proposed method. These videos have been manipulated with four different state-of-the-art facial manipulation methods: DeepFake,[1] FaceSwap,[2] Face2Face [33] and NeuralTextures [32], each of them consists of three kinds of image qualities: RAW, HQ, LQ, which respectively stands for the original video quality, high quality, and low quality. In our experiment, we took the FaceForensics++ dataset for the training and testing task.

[1] www.github.com/deepfakes/faceswap.

[2] www.github.com/MarekKowalski/FaceSwap.

Table 1 Comparison with different network backbones on the FaceForensics++ dataset

Network backbones	Classification accuracy		
	LQ (%)	HQ (%)	RAW (%)
AlexNet [13]	86.89	88.64	91.97
VGG [28]	90.85	92.35	94.67
ResNet [9]	93.42	95.68	96.43
Efficient Net(Ours)	96.51	97.89	99.57

Table 2 Comparison with different methods on the FaceForensics++ dataset

Methods	Classification accuracy		
	LQ (%)	HQ (%)	RAW (%)
CNN [23]	90.00	91.45	93.40
SVM [19]	70.10	73.64	75.43
RNN [24]	93.46	95.04	95.98
GRU [6]	94.48	96.18	97.54
LSTM [3]	94.29	96.24	96.79
ConvLSTM [27]	95.18	96.79	98.80
ConvLstm (with attention)	96.51	97.89	99.57

In addition, we also augmented the input facial data to evaluate the performance of our approach. The augmented inputs consist of the Gaussian blended videos and the videos in which the nose, eyes, and mouth were separately removed from the original videos. For the implementation detail, we separated the dataset as the ratio of 80:10:10 for the training, testing, and validation task. In order to improve the efficiency of our network, we took 15 random frames of each video as the input, which were all converted from different resolutions into the same size of 224×224.

4.2 Implementation Details

For CNN, we chose the EfficientNet-B5 as the backbone of our network. Table 1 shows the experiment on the comparison between different network backbones on the FaceForensics++(LQ) dataset.

As is shown in Table 1, compared to the traditional CNN-based backbone AlexNet, the EfficientNet has achieved a classification accuracy promotion of 10% followed by the same network. The computer used for the experiments was equipped with two NVIDIA 1080Ti GPUs, 32 GB RAM and an INTEL® Xeon(R) E5-2609 V4 @1.70 GHz $\times 8$ CPU. We utilize the PyTorch platform to make all experiments.

4.3 Experiment on the Effectiveness of Convolutional LSTM

One of the contributions of our work is the application of the attention-based ConvLSTM model in our network, which can effectively utilize the facial feature of the input frames. Such a design can effectively improve the performance of the detector. We compared our approach with some classic DeepFake detection methods [1,3,24,49] to evaluate the performance of our network, the experimental results are shown in Table 2.

Table 3 Experimental results of different attention model on the FaceForensics++ dataset

Methods	Classification accuracy		
	LQ (%)	HQ (%)	RAW (%)
ConvLSTM	95.18	96.79	98.80
ConvLSTM+hard attention	95.22	96.91	99.24
self-attention	95.96	97.34	98.91
ConvLSTM+soft attention	96.51	97.89	99.57

Through these experiments, we can find out that our method achieves a more significant performance promotion in comparison with other methods. Some observations are as follows:

- Mesonet [1] focuses on the mesoscopic properties of images using a low number of layers, whereas the correlation across video frames is ignored. In other words, it does not consider the facial information crossing frames in the training step. Such a method may achieve a reasonable result for the task of image detection. However, the result proved to be not very well in terms of the DeepFake video detection where the facial information crossing frames information is an essential factor that needs to be considered.
- RNN [24] considers the correlation of the temporal feature in DeepFake videos. As a result, it performances better than the Mesonet, but it still cannot match other latest methods. For the reason that the RNN addresses the temporal information from the related input, it can find out the relation. However, it has the drawback of discarding the previous inputs and failing to describe the correlations of these input video frames, which leads to the overlap of the videos. All these issues bring redundant information and thus reduce the accuracy of the classifier.
- The LSTM-based model [3] performs better than Mesonet. Without a doubt, such architecture takes temporal information into consideration. The pre-trained CNN [5] is applied to extract the facial feature from the inputs. In other words, the pre-trained CNN model can extract robust feature vectors for the input DeepFake videos. As a result, it will make the model consider the cross-frame facial feature information of the feature vector.
- GRU is the simplified version of the LSTM model where its parameters are effectively reduced while maintaining its performance. The traditional LSTM model is endowed with lots of parameters, as a result, training is a difficult task. By packing the input gate and the forget gate into the update gate, the GRU can provide a significant improvement of the training speed and reduce the computing cost.
- Our approach applied ConvLSTM [27] to obtain both temporal and spatial information of the input sequential data. As a result, the ConvLSTM is very suitable for the task of processing both the inter-frame and the intra-frame features crossing the input frames.

4.4 Experiment on the Effectiveness of the Soft-Attention Model

In this section, further experiments were conducted to analyze the effectiveness of the soft-attention weight mechanism. As mentioned before, the attention mechanism was applied to remove redundant information.

We tested the effectiveness of different attention-based feature aggregation models with a series of experiments. The results are shown in Table 3.

We can make some observations as follows:

Table 4 Experiment on the effectiveness of video frame extraction methods

Sampling methods	Classification accuracy (%)		
	LQ	HQ	RAW
Equally Spaced [29]	95.74	96.94	98.14
Key Frame [20]	96.09	97.42	98.95
Sub-sequence [8]	95.38	96.43	97.83
Sequential [21]	94.63	95.48	96.98
Random (Ours)	96.51 ± 0.38	97.89 ± 0.34	99.57 ± 0.29

- The hard-attention mechanism [40] has the ability to focus on one exact position of the input at a time. It enables the network to focus on relatively important information. However, the hard-attention mechanism cannot take care of all positions of the input and is not differentiable, this may cause performance decrease in comparison with other attention-based methods.
- Self-attention [35] solves the problem of parallel computing in the RNN model. It can be seen as a network based on the encoding-decoding structure. Both the encoder and the decoder are based on the multi-head structure. Such a design eliminates duplication and convolution of the RNN-based models, thus significantly save the computing cost. However, it might have the problem of ignoring the temporal features of the inputs.
- We applied the soft-attention mechanism to our model that combines the CNN and ConvLSTM, which could be seen as the video feature processing part of our method.
  Since the visual feature of the video frames has a lot of redundant information, the attention mechanism we have mentioned cannot remove it thoroughly. We can observe a significant performance improvement with the employment of the soft-attention mechanism. Such a mechanism can guide the network to focus more on the significant feature while maintaining the integrity of the network input.

4.5 Experiment on the Effectiveness of Video Frame Extraction Methods

In order to verify the correlation between the classification performance and the video frame extraction methods, we compared our methods with different frame sampling methods. All the input frames are set to 15 as a control variable, the results are shown in Table 4.

Results show that when the network structure remains constant, our sampling method achieves the best classification accuracy. The sequential sampling method applied by Nguyen et al. [21] did not strengthen the incoherence of facial data across frames and thus had the worst performance. Moreover, the sub-sequence method proposed by Guera et al. [8] sliced the input video into sequential frames, and its performance has improved significantly. However, it still suffers from the identical drawback. The equally spaced method applied by Singh et al. [29] improved the classification accuracy by strengthening the incoherence of facial data across frames, whereas it has the drawback of bringing redundant information. The key frame extraction method proposed by Mitra et al. [20] has the problem of drastic performance fluctuation when the scene changes sharply. In comparison with these methods, our method has taken the incoherence of facial data across frames into account while reduced the redundancy, and thus, has the best classification accuracy.

Table 5 Comparisons of classification experimental results on the FaceForensics++(LQ) dataset

Study	Method	Classifiers	Classification accuracy (%)	Dataset
Afchar et al. [1]	Mesoscopic features	CNN	83.2	F2F
Rössler et al. [23]	Steganalysis features	CNN	81	NT
			91	F2F
			94	DF
			93	FS
Sabir et al. [24]	Temporal features	RNN	94.3	F2F
Amerini et al. [4]	Temporal features	OpticalFlow	81.6	F2F
Wang et al. [37]	Deep learning features	3DCNN	95.1	DF
			92.3	FS
Ours	Interframe features	ConvLSTM	**96.70**	DF
			96.51	F2F
			94.85	FS
			92.71	NT

Bold values highlight the best performance in the comparison of the state-of-the-art methods

Table 6 Experimental results of the variation of facial data on the FaceForensics++ dataset

Methods	Classification accuracy (%)		
	LQ	HQ	RAW
Original	94.21	95.01	97.39
Gussian blur	94.38	95.76	97.43
Wavelet transform	94.96	95.87	97.98
Facial data removal	96.51	97.89	99.57

4.6 Experiment on the Variation of Facial Data

For the majority of the DeepFake video detection network, there is always a significant performance increment when the diversity of the input faces increases. Some features like the eye blinking, nose movement and the details around the lips pose are also inconsistent in deepfake videos, which is also an essential element in the detection tasks. In order to enable our network the ability of learning more facial features, we separately removed the facial elements including eyes, nose, and mouth in Deepfake videos. Inspired by the methods taken in [16] we took the Gaussian Blur of the located face area as the variation of the input. We also took the wavelet transform method to improve classification performance under low-quality images.

Our experimental results are presented in Table 6. Results show that the classification accuracy increases when the resolution changes of the manipulated area are strengthened through either the Gaussian blur or the wavelet transform in comparison with the original inputs. The classification performance increases significantly when the particular feature of the facial data is strengthened through the removal of other features, which demonstrates the efficiency of particular facial features in the task of DeepFake detection.

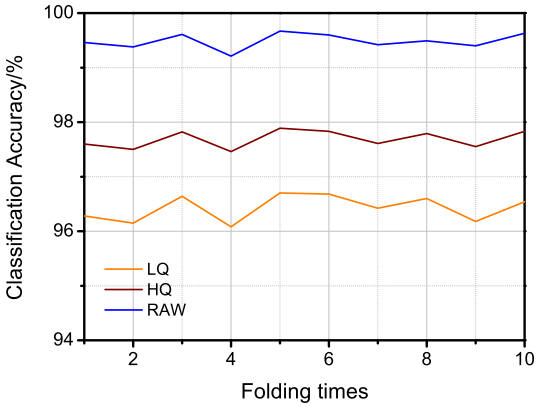

Fig. 3 Classification accuracy using 10-fold cross validation, note that the horizontal axis stands for the times when different folds are chosen for training and testing

4.7 Comparison With State-Of-The-Art Methods on the FaceForensics++ Dataset

To validate the efficiency of our network, our proposed method was compared with the state-of-the-art methods on the FaceForensics++ dataset. These methods include the CNN-based methods [1,16,23], the RNN-based methods [24], the visual artifacts-based methods [41] the Multi-Layer Perceptron(MLP)-based method [18]. The results are shown in Table 5.

The results show that our detection method has achieved the best accuracy. The traditional computer vision-based methods, however, have the worst classification accuracy. The CNN-based methods applied deep learning to extract the manipulated feature of DeepFake videos, and its performance has significantly improved. Similarly, there is also an improvement in the performance when the RNN-based network is applied to enforce the temporal coherence across the video frames. In comparison with these methods, our method has the advantage of applying weights into the input videos and taking advantage of the facial feature incoherence, which promotes both classification accuracy and processing efficiency.

4.8 Cross Validation

The existing detection methods based on the FaceForensics++ dataset mainly split the data into a certain ratio for training and testing tasks. Such a method may have the drawback of overfitting due to inadequate data splitting methods. The adverse effect caused by unbalanced data splitting methods in a single division can be minimized by using cross validation methods. To better evaluate the generalization performance of our model, we use 10-fold cross validation to separate training and testing sets on the FaceForensics++ dataset. The experimental results are shown in Fig. 3.

Experimental results demonstrate that our detection method achieves a high classification accuracy while maintaining satisfactory generalization ability.

4.9 Complexity Analysis

To evaluate the computational complexity, we conducted comparative experiments with some typical methods on the MI3DOR benchmark. The experimental results for the size of the parameter and the computation time are shown in Table 7. Here, the running time represents the time of the current method getting a steady performance, and the testing time represents the

Table 7 Comparison on the computational complexity

Methods	Model size (MB)	Running time (s)	Testing time (ms)
McCloskey et al. [19]	59	793.5	15.5
Galteri et al. [1]	60	1248.4	16.9
Rössler et al. [23]	66	1983.8	18.6
Sabir et al. [24]	63	2099.4	19.7
Ours	61	1731.8	18.4

forward pass time. An examination of the results shows that our method achieves reasonable computational efficiency while our method reaches a steady performance in few iterations. McCloskey et al. employed the traditional machine learning-based method, so the computational complexity is great, but the retrieval performance is not well. The rest methods utilized the inter-frame features to improve the classification performance, whereas the convolution operation takes most of the training time. Although our method employed the ConvLSTM and the attention mechanism, the algorithm complexity is less than the convolution operation. So our method reaches reasonable computational efficiency.

5 Conclusion

This paper presents a novel attention-based LSTM network to detect DeepFake videos with high efficiency and low computational cost. Specifically, the soft-attention mechanism based on weights defined by the visual saliency model is applied to save both the visual information and the correlation information of the input stream in the training process. Compared with current effective methods, not only does our method utilize the visual information from the input videos, but it takes the correlation information into concern as well. Experimental results on the public dataset demonstrate the effectiveness of the proposed network, which means the correlation information is crucial for DeepFake video detection methods. Moreover, the experimental results around the augment facial dataset demonstrate that the eyes and mouth play a paramount role in the detection of Deepfake videos.

Acknowledgements This work was supported in part by the National Key Research and Development Program of China (2020YFB1711704) and the National Natural Science Foundation of China (61872267, 61772359, 61572356, 61862020, 61861014).

References

1. Afchar D, Nozick V, Yamagishi J, Echizen I. Mesonet: a compact facial video forgery detection network. In: 2018 IEEE international workshop on information forensics and security (WIFS), Hong Kong, China, December 11–13. IEEE, pp 1–7
2. Fawad A, Mohammed Yakoob S, Vali Uddin A (2010) A secure and robust hash-based scheme for image authentication. Signal Process 90(5):1456–1470
3. Amerini I, Caldelli R (2020) Exploiting prediction error inconsistencies through lstm-based classifiers to detect deepfake videos. In: Christian R, Franziska S, Irene A, Paolo B, Tomás P (eds) IH&MMSec '20: ACM workshop on information hiding and multimedia security, Denver, CO, USA, June 22–24. ACM, pp 97–102

4. Amerini I, Galteri L, Caldelli R, Del Bimbo A (2019) Deepfake video detection through optical flow based CNN. In: 2019 IEEE/CVF international conference on computer vision workshops, ICCV workshops 2019, Seoul, Korea (South), October 27–28. IEEE, pp 1205–1207
5. Amerini I, Li C-T, Caldelli R (2019) Social network identification through image classification with CNN. IEEE Access 7:35264–35273
6. Chung J, Gulcehre C, Cho K, Bengio Y (2014) Empirical evaluation of gated recurrent neural networks on sequence modeling. CoRR, abs/1412.3555
7. Cornia M, Baraldi L, Serra G, Cucchiara R (2018) Predicting human eye fixations via an LSTM-based saliency attentive model. IEEE Trans Image Process 27(10):5142–5154
8. Güera D, Delp EJ (2018) Deepfake video detection using recurrent neural networks. In: 15th IEEE international conference on advanced video and signal based surveillance, AVSS 2018, Auckland, New Zealand, November 27–30. IEEE, pp 1–6
9. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: 2016 IEEE conference on computer vision and pattern recognition, CVPR 2016, Las Vegas, NV, USA, June 27–30. IEEE Computer Society, pp 770–778
10. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780
11. Hong C, Jun Yu, Zhang J, Jin X, Lee K-H (2019) Multimodal face-pose estimation with multitask manifold deep learning. IEEE Trans Ind Inform 15(7):3952–3961
12. Kalchbrenner N, Danihelka I, Graves A (2016) Grid long short-term memory. In: Bengio Y, LeCun Y (eds) 4th international conference on learning representations, ICLR 2016, San Juan, Puerto Rico, May 2–4, 2016, conference track proceedings
13. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In Bartlett PL, Pereira FCN, Burges CJS, Bottou L, Weinberger KQ (eds) Advances in neural information processing systems 25: 26th annual conference on neural information processing systems 2012. Proceedings of a meeting held December 3–6, 2012, Lake Tahoe, Nevada, United States, pp 1106–1114
14. Li X, Zhang W, Ding Q (2019) Understanding and improving deep learning-based rolling bearing fault diagnosis with attention mechanism. Signal Process 161:136–154
15. Li Y, Chang M-C, Lyu S (2018) In ICTU oculi: exposing AI created fake videos by detecting eye blinking. In: 2018 IEEE international workshop on information forensics and security, WIFS 2018, Hong Kong, China, December 11–13. IEEE, pp 1–7
16. Li Y, Lyu S (2018) Exposing deepfake videos by detecting face warping artifacts. volume abs/1811.00656
17. Ma Y, Peng H, Cambria E (2018) Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive LSTM. In: McIlraith SA, Weinberger KQ (eds) Proceedings of the thirty-second AAAI conference on artificial intelligence (AAAI-18), the 30th innovative applications of artificial intelligence (IAAI-18), and the 8th AAAI symposium on educational advances in artificial intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2–7, 2018. AAAI Press, pp 5876–5883
18. Matern F, Riess C, Stamminger M (2019) Exploiting visual artifacts to expose deepfakes and face manipulations. In: 2019 IEEE winter applications of computer vision workshops (WACVW), pp 83–92
19. McCloskey S, Albright M (2018) Detecting Gan-generated imagery using color cues. CoRR, abs/1812.08247
20. Mitra A, Mohanty SP, Corcoran P, Kougianos E (2021) A machine learning based approach for deepfake detection in social media through key video frame extraction. SN Comput Sci 2(2):98
21. Nguyen HH, Fang F, Yamagishi J, Echizen I (2019) Multi-task learning for detecting and segmenting manipulated facial images and videos. In: 10th IEEE international conference on biometrics theory, applications and systems, BTAS 2019, Tampa, FL, USA, September 23–26. IEEE, pp 1–8
22. Park M (2020) JGAN: a joint formulation of GAN for synthesizing images and labels. IEEE Access 8:188883–188888
23. Rössler A, Cozzolino D, Verdoliva L, Riess C, Thies J, Nießner M (2019) Faceforensics++: learning to detect manipulated facial images. In: 2019 IEEE/CVF international conference on computer vision, ICCV 2019, Seoul, Korea (South), October 27–November 2. IEEE, pp 1–11
24. Sabir E, Cheng J, Jaiswal A, AbdAlmageed W, Masi I, Natarajan P (2019) Recurrent convolutional strategies for face manipulation detection in videos. In: IEEE conference on computer vision and pattern recognition workshops, CVPR workshops 2019, Long Beach, CA, USA, June 16–20, 2019. Computer vision foundation/IEEE, pp 80–87
25. Seelamantula CS, Sreenivas TV (2009) Blocking artifacts in speech/audio: dynamic auditory model-based characterization and optimal time-frequency smoothing. Signal Process 89(4):523–531
26. Shalaby MAW, Ahmad MO (2013) A multilevel structural technique for fingerprint representation and matching. Signal Process 93(1):56–69

27. Shi X, Chen Z, Wang H, Yeung D-Y, Wong W-K, Woo W-c (2015) Convolutional LSTM network: a machine learning approach for precipitation nowcasting. In: Cortes C, Lawrence ND, Lee DN, Sugiyama M, Garnett R (eds) Advances in neural information processing systems 28: annual conference on neural information processing systems 2015, December 7–12, Montreal, Quebec, Canada, pp 802–810
28. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. In: Bengio Y, LeCun Y (eds) 3rd international conference on learning representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, conference track proceedings
29. Singh A, Saimbhi AS, Singh N, Mittal M (2020) Deepfake video detection: a time-distributed approach. SN Comput Sci 1(4):212
30. Tan M, Le QV (2019) Efficientnet: rethinking model scaling for convolutional neural networks. In: Chaudhuri K, Salakhutdinov R (eds) Proceedings of the 36th international conference on machine learning, ICML 2019, 9–15 June 2019, Long Beach, CA, USA, volume 97 of Proceedings of machine learning research. PMLR, pp 6105–6114
31. Thies J, Zollhöfer M, Stamminger M, Theobalt C, Nießner M (2016) Face2face: Real-time face capture and reenactment of RGB videos. In: 2016 IEEE conference on computer vision and pattern recognition (cVPR), pp 2387–2395
32. Thies J, Zollhöfer M, Nießner M (2019) Deferred neural rendering: image synthesis using neural textures. ACM Trans Graph 38(4):66:1-66:12
33. Thies J, Zollhöfer M, Stamminger M, Theobalt C, Nießner M (2020) Face2face: real-time face capture and reenactment of RGB videos. CoRR, abs/2007.14808
34. Tolosana R, Vera-Rodríguez R, Fiérrez J, Morales A, Ortega-Garcia J (2020) Deepfakes and beyond: a survey of face manipulation and fake detection. Inf Fusion 64:131–148
35. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. In: Guyon I, von Luxburg U, Bengio S, Wallach HM, Fergus R, Vishwanathan SVN, Garnett R (eds) Advances in neural information processing systems 30: annual conference on neural information processing systems 2017, 4–9 December 2017, Long Beach, CA, USA, pp 5998–6008
36. Wang N, Zhang S, Gao X, Li J, Song B, Li Z (2017) Unified framework for face sketch synthesis. Signal Process 130:1–11
37. Wang Y, Bilinski P, Brémond F, Dantcheva A (2020) G3AN: disentangling appearance and motion for video generation. In: 2020 IEEE/CVF conference on computer vision and pattern recognition, CVPR 2020, Seattle, WA, USA, June 13–19. IEEE, pp 5263–5272
38. Wang Y, Wang S, Qi G, Tang J, Li B (2018) Weakly supervised facial attribute manipulation via deep adversarial network. In: 2018 IEEE winter conference on applications of computer vision, WACV 2018, Lake Tahoe, NV, USA, March 12–15. IEEE Computer Society, pp 112–121
39. Shaoen W, Junhong X, Zhu S, Guo H (2018) A deep residual convolutional neural network for facial keypoint detection with missing labels. Signal Process 144:384–391
40. Xu K, Ba J, Kiros R, Cho K, Courville AC, Salakhutdinov R, Zemel RS, Bengio Y (2015) Show, attend and tell: Neural image caption generation with visual attention. In: Bach FR, Blei DM (eds) Proceedings of the 32nd international conference on machine learning, ICML 2015, Lille, France, 6–11 July 2015, volume 37 of JMLR workshop and conference proceedings. JMLR.org, pp 2048–2057
41. Yang X, Li Y, Lyu S (2019) Exposing deep fakes using inconsistent head poses. In: IEEE international conference on acoustics, speech and signal processing, ICASSP 2019, Brighton, UK, May 12–17, 2019. IEEE, pp 8261–8265
42. Yang Y, Zhou J, Ai J, Bin Y, Hanjalic A, Shen HT, Ji Y (2018) Video captioning by adversarial LSTM. IEEE Trans Image Process 27(11):5600–5611
43. Jun Yu, Rui Y, Tao D (2014) Click prediction for web image reranking using multimodal sparse coding. IEEE Trans Image Process 23(5):2019–2032
44. Yu J, Tan M, Zhang H, Tao D, Rui Y (2019) Hierarchical deep click feature prediction for fine-grained image recognition. IEEE Trans Pattern Anal Mach Intell https://doi.org/10.1109/TPAMI.2019.2932058
45. Jun Yu, Tao D, Wang M, Rui Y (2015) Learning to rank using user clicks and visual features for image retrieval. IEEE Trans Cybern 45(4):767–779
46. Zhang H, Goodfellow IJ, Metaxas DN, Odena A (2019) Self-attention generative adversarial networks. In: Chaudhuri K, Salakhutdinov R (eds) Proceedings of the 36th international conference on machine learning, ICML 2019, 9–15 June 2019, Long Beach, CA, USA, volume 97 of Proceedings of machine learning research. PMLR, pp 7354–7363
47. Zhang K, Zhang Z, Li Z, Qiao Yu (2016) Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process Lett 23(10):1499–1503
48. Zhang Y, Zheng L, Thing VLL (2017) Automated face swapping and its detection. In: 2017 IEEE 2nd international conference on signal and image processing (ICSIP), pp 15–19

49. Zhou P, Han X, Morariu VI, Davis LS (2017) Two-stream neural networks for tampered face detection. In: 2017 IEEE conference on computer vision and pattern recognition workshops, CVPR workshops 2017, Honolulu, HI, USA, July 21–26. IEEE Computer Society, pp 1831–1839
50. Zhu J-Y, Park T, Isola P, Efros AA (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. In: IEEE international conference on computer vision, ICCV 2017, Venice, Italy, October 22–29. IEEE Computer Society, pp 2242–2251
51. Barret Z, Le QV (2017) Neural architecture search with reinforcement learning. In: 5th international conference on learning representations, ICLR 2017, Toulon, France, April 24–26, 2017, conference track proceedings. OpenReview.net