Pattern Analysis and Applications (2022) 25:981–992
https://doi.org/10.1007/s10044-022-01083-2

SHORT PAPER

Visual attention-based deepfake video forgery detection

Shreyan Ganguly[1] · Sk Mohiuddin[2] · Samir Malakar[2] · Erik Cuevas[3] · Ram Sarkar[4]

Received: 3 November 2021 / Accepted: 19 May 2022 / Published online: 27 June 2022

Abstract
The prime goal of creating synthetic digital data is to generate something very closer to real ones when the original data are scarce. However, the trustworthiness of such digital content is dipping potentially in society owing to malicious users. Deepfake method that uses computer graphics and computer vision techniques to replace the face of one person with the face of a different person is becoming an area of big concern. Such techniques can easily be used to hide the identity of a person. Therefore, a method is needed to verify the originality of such face images/videos. To this end, we design a deep learning model enhanced with visual attention technique to differentiate manipulated videos/images (generated by deepfake methods) from real ones. At first, we extract the face region from video frames and then pass the same through the pre-trained Xception model to obtain the feature maps. Next, with the help of the visual attention mechanism, we mainly try to focus on the deepfake video manipulation leftover artifacts. We evaluate our model on two publicly available datasets, namely FaceForensics++ and Celeb-DF (V2), and our model outperforms many state-of-the-art methods tested on these two datasets. Source code of the proposed method can be found at: https://github.com/tre3x/Deepfake-Video-Forgery-Detection.

Keywords Deepfake · Visual attention · Video forgery · Deep learning · FaceForensics++ · Celeb-DF (V2)

1 Introduction

The advent of generative adversarial networks (GANs) [1] has opened up many new research avenues in the computer vision domain since the last few years. Undoubtedly, the most important task tackled by GANs is to generate new synthetic samples from an existing collection of samples where the original samples are scarce. The goal of these GANs is to generate new unseen data, mostly images, from scratch, which is possible by extensive training of generative and discriminative models on the original set of data. Some of the important applications of GANs include cartoon character generation, clothing translation, high-quality image generation, etc. However, this synthetic data generation has some negative implications for society that raises some security issues. For example, one can easily generate an almost realistic but artificial image/video by manipulating the target image/video with the help of GANs. Such scenarios create a new research problem which is the detection of a manipulated image/video produced by some means using machine learning or deep learning models.

The continuous evolution of GANs has resulted in the generation of high-quality digital data with minimal observable error, which makes authentication of these synthetic

✉ Erik Cuevas
erik.cuevas@cucei.udg.mx

Shreyan Ganguly
gshreyan16@gmail.com

Sk Mohiuddin
myselfmohiuddin@gmail.com

Samir Malakar
malakarsamir@gmail.com

Ram Sarkar
raamsarkar@gmail.com

[1] Department of Construction Engineering, Jadavpur University Salt Lake Campus, Kolkata, West Bengal 700098, India

[2] Department of Computer Science, Asutosh College, Kolkata, West Bengal 700026, India

[3] Departamento de Electrónica, Universidad de Guadalajara, 44430 City, C.P, México

[4] Department of Computer Science and Engineering, Jadavpur University, Kolkata, West Bengal 700032, India

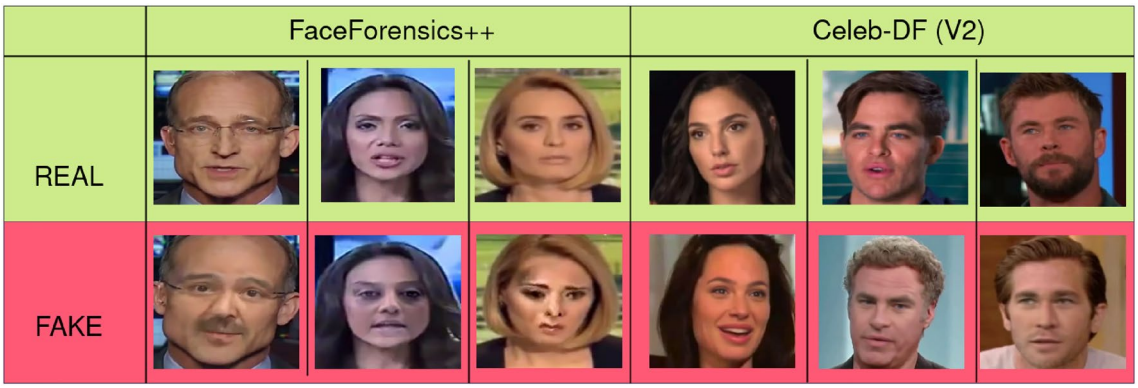


Fig. 1 Some pairs (column-wise) of cropped face images taken from real and fake videos of FaceForensics++ [11] and Celeb-DF (V2) [12] datasets

data more difficult. The process of face manipulation has also been evolving, and the detection of such manipulation needs to follow the former's pace. Detection of face manipulated images or videos generated by some GAN models mostly relies on the leftover artifacts generated by the underlying GAN model. The early days' works have mainly focused on retina color [2], eye blinking pattern [3], different facial regions [4], and a few are based upon the visual artifacts left by the GAN-based manipulator [5, 6]. Most of the recent works used some convolutional neural network (CNN)-based models to detect such manipulations. For example, pre-trained Xception [7] and Capsule Network [8] are used as a backbone for developing such detection models. Recently, Tolosana et al. [4] have considered two different inputs from a frame for detecting fake videos manipulated by some deepfake method. One input is the entire face image, while the other one is the specific facial region extracted by segmenting the input facial image. This work has shown that the use of information from some specific regions helps in improving the overall performance of the model.

It has been observed that there are some visible artifacts in the synthetically generated video frames (see Fig. 1). Hence, designing a deep-learning-based feature extractor which tries to look for the presence or absence of leftover artifacts can be a viable technique to detect deepfakes. Recently, the Xception network [7] has gained a lot of attention in the domain of deepfake image detection [9, 10]. However, it should also be noted that some features generated by this network may be redundant for decision making. Hence, giving less importance to the redundant ones can be considered as a crucial step to improve the model performance. A lot of recent deep learning approaches in prioritizing important features use attention mechanisms, which are mostly computationally expensive. Keeping these facts in mind, in this work, we propose a light-weight soft attention mechanism on top of the Xception network. In doing so, at first, we extract specific frames from videos and then crop the face region from the frames. The cropped face region is used to detect whether the given video is manipulated or not. Features are extracted from cropped faces by the Xception model and top of which we apply the soft attention mechanism. In this way, we have tried to pass the more relevant features like the artifacts left by the synthetic face image generation method to our classification model. These artifacts help to detect the authenticity of the questioned image/video efficiently. Upon extensive experimentation on two popular and challenging datasets, we try to make sure that the soft attention mechanism is useful to provide more priority to important features and helps in making better decisions whether a facial image present in video is real or fake.

In a nutshell, the highlights of our work are as follows:

- We have designed a novel deep-learning-based deepfake detection system which tries to make decision based on specific features generated by a CNN model signifying the presence or absence of synthetic artifacts.
- We have used a light-weight attention mechanism technique that helps in prioritizing the features useful in making better decisions.
- We have validated our system on two popular and challenging benchmark datasets - FaceForensics++, Celeb-DF (V2), and it is observed that the system outperforms many state-of-the-art methods.

The remaining part of this article is organized as follows. Section 2 describes previous works followed by the motivation in Sect. 3. Our proposed method is described in Sect. 4, while Sect. 5 is used for results and discussion. Finally, the paper is concluded in Sect. 6.

2 Previous work

It has been observed that the first generation of video/image content change in terms of fake faces possesses ample variation concerning real ones. Hence, detection approaches for such videos/images have mainly considered the inconsistencies of pixel properties. Later, in the second generation, it has been observed that GAN-based forged data are very similar to the original ones and such data come with higher resolution. However, for authentication of recent face manipulated videos/images, the performances of deep learning-based models are proved to be much superior to traditional machine learning or basic image processing approaches. Below we describe some past methods which deal with the problem under consideration.

2.1 Hand-crafted feature-based deepfake detection methods

Deepfake detection techniques using hand-crafted features are straightforward and have advantages like less time to detect and can work on low-computational resources. Thus, several researchers in past have tried to deal with this problem using hand-engineered features and classical machine learning. For example, Durall et al. [13] have analyzed the unnatural behavior of synthesis video using discrete Fourier transform (DFT) and converted them from spatial domain to 1D power spectrum which is then fed to classical classifiers like logistic regression and support vector machine (SVM) for authentication. A similar method is proposed by Guarnera et al. [14] that uses the expectation maximization (EM) algorithm [15] on each channel of the input image to extract pixel correlation and classifies face images using different naive classifiers. Yang et al. [16] have proposed a method to expose fake faces by estimating head poses using 68 facial landmarks from central face regions. They have shown that the position of facial landmarks is shifted during tampering. Li et al. [17] have designed a method that uses the disparity observed in H, S, Cb, and Cr color components of the fake images. It first extracts four color channels (H, S, Cb, Cr) from the input images and then converts each of them into four co-occurrence matrices. Finally, element-wise sums are taken as features to detect the forgery. The camera-captured noises are used by Koopman et al. [18] in their work. Equal-sized cropped faces from videos are sequentially grouped and measured photo response non-uniformity (PRNU) pattern for each group. Average correlation and variations in correlation scores among the groups are passed through Welch's t-test [19] to identify manipulated faces. These methods have performed well on the first generation of synthetic videos/images. However, these are not capable of tracking all kinds of inconsistencies present in the second-generation manipulated images/videos.

2.2 Deep learning-based deepfake detection methods

Recently, deepfake detection using deep learning-based methods has become widespread due to their learnable features that produce a better performance as compared to its hand-crafted counterparts. Most of the methods [4, 20, 21] have used the transfer learning protocol. Shang et al. [22] have built a CNN architecture that uses two modules stacked one after another. The first module captures pixel relation with their nearest one and the other one focuses on the region relation among original with the manipulated ones. They use a high-resolution network (HRNet-w30) [23] at the beginning to generate images with different resolution levels for a single input image and then feed these newly generated images to the second module. To focus on mesoscopic properties of an image, Afchar et al. [20] have designed a model composed of two different CNN architectures: i) Meso-4, a CNN network composed of 4 convolutional layers followed by a fully connected layer, and ii) MesoInception-4, replaced the first two layers of Meso-4 by the variant of Inception module [24]. An exhaustive study is carried out by Rossler et al. [21] using the FaceForensics++ benchmark dataset prepared by them. The authors have shown that XceptionNet [7] performs the best among others. In another work [25], Bayar and Stamm have designed their own CNN model to adaptively learn editing features from images. This method is designed to detect universal image manipulations rather than detecting a specific type of manipulation. The unnatural motion of facial components is brought into notice by Amerini et al. [26] where a CNN-based model PWC-Net [27] is used to extract optical flow matrices from the face images and then feed these into VGG16 [28]-based semi-trainable model to filter out authentic videos.

Recent studies show that researchers use attention mechanisms to focus on certain parts of input images which are useful for classification. However, the attention approach widely used for different purposes [29, 30] is computationally expensive. Su et al. [31] used convolutional long

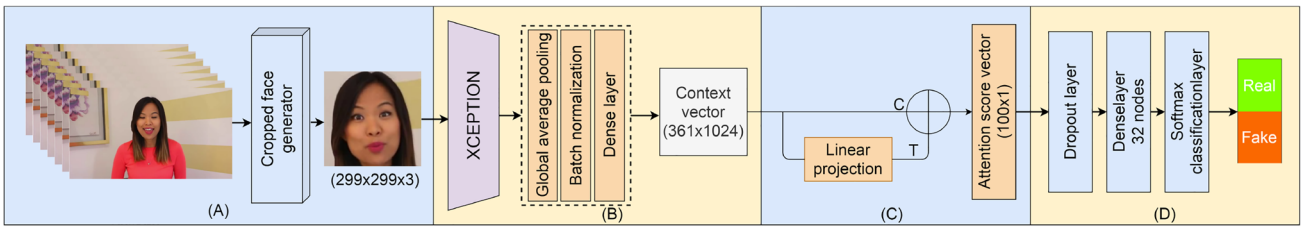


Fig. 2 Proposed visual attention-based deepfake detection architecture. **A** represents cropped face generator which involves frame extraction and detection, and cropping of the face from extracted frames, **B** shows the generation of feature maps using the Xception model from cropped face images. **C** illustrates the working process of the soft attention network that is applied on feature maps. **D** represents the classification network that consists of two dense layers and a softmax layer. Using Xception as backbone CNN, we get the dimension of the feature map generated as $(19 \times 19 \times 1024)$, and the context vector dimension as (361×1024). This context vector passes to the soft attention network

short-term memory (LSTM) to extract both spatial and temporal information of the deepfake video. Further, soft attention was applied to the output of CNN features to identify deepfake videos. Khormali and Yuan [32] considered zooming in and zooming out face images to extract separate feature elements and then paid attention to them for deciding the authenticity of the video. An attention-based data augmentation method was proposed by Wang and Deng [33] which guided the detector to refine and enlarge its focus to identify face forgeries. To exploit local features of the face region, an attention mechanism was applied by Chen et al. [34] where information represented by RGB and frequency domain was fused to extend the model's capability of artifact pattern similarity learning.

3 Motivation

Visual multimedia files are widely perceived as offering authentic evidence of actual events, including, in particular, the presence and actions of subjects (mainly humans) in images and videos. Although this perception is slowly shifting, contemporary technologies allow far easier and more accessible manipulation of these media files. This gap represents a societal threat whenever manipulated media are released over social networks and consumed by the public that is ill-equipped to question its genuineness. As a result, the detection of video forgery is becoming a pressing need. Synthetically generated fake videos look inseparable from original videos in the naked eyes; however, there are subtle differences between them such as visible artifacts in the target face or whole frame of video in general as shown in Fig. 1. In our work, we exploit the presence of these synthetic artifacts in the video frames for the detection of deepfake videos. A lot of past methods, as discussed in Sect. 2, try to make decisions based on the presence of these fake artifacts by using attention-based techniques. However, most of the attention-based techniques used are computationally expensive. In this work, we use a light-weight soft attention-based technique which tries to put more focus on important features generated by a CNN model. The proposed method achieves state-of-the-art results on two benchmark datasets.

4 Proposed Method

In this section, we describe the proposed method which distinguishes deepfake videos or images from real ones. The proposed model extracts frames from videos and generates cropped faces which are used to detect manipulated videos. Feature maps from cropped faces are generated using the Xception model. With the help of the attention mechanism applied on the feature maps, we mainly focus on highlighting the artifacts on the questioned video frame that can help to detect the authenticity of the images or videos. The pipeline of the entire method is shown in Fig. 2 pictorially, while Algorithm 1 shows the steps that we followed in this research. In the following subsections, we describe each module of the proposed system.

4.1 Data preparation

In this stage, we extract frames from a video and crop out the face region from the extracted frames. The entire process is illustrated in Fig. 3. We begin by selecting N random moderately compressed P-frames from the video. The value of N depends on the experimental setup.

Algorithm 1 Deepfake Detection Method

Input: Videos for authentication.
Output: Classification of the videos as Real or Fake.
Step 1: Extract N random frames from the input video.
Step 2: Detect face region inside a frame using Multi-Task Cascading Neural Network (MTCNN) model [35].
Step 3: Increase the detected face region by 10% in each direction if the extended dimension is viable and crop the face.
Step 4: Resize the cropped face image into dimension (299×299) pixels.
Step 5: Feed the face image into a sliced Xception module to generate the feature maps.
Step 6: Apply global average pooling and batch normalization on generated feature maps to transform into dimension $19 \times 19 \times 1024$.
Step 7: Employ affine transformation using dense layers on the generated feature maps to generate a transformed feature map.
Step 8: Highlight the relevant features using a soft attention mechanism to discriminate properties of a synthetic face.
Step 9: Pass the output of Step 8 through two dense layers followed by a softmax classification layer to classify whether a frame image is real or fake.

Next, we pass the selected frames to Multi-Task Cascading Neural Network (MTCNN) [35]-based face detection algorithm to detect the face present in the frame. We use this network because of its lightweight architecture which uses a multi-task learning mechanism and hence works well with faces occupying different sizes in the image. The used method also performs better in different lighting conditions. After detecting the bounding box around the face region using MTCNN, we increase the size of the detected bounding box by 10% to include some background information around the face region. Crop faces are resized to (299×299) dimension which is the default input resolution of the Xception model [7]. It is worth mentioning that the Xception model is used here to extract feature maps from the cropped face images.

Fig. 3 Illustration of image data preparation from a video

4.2 Generation of feature maps using Xception model

Nowadays, it becomes a common practice to use a pre-trained CNN model as a fixed feature extractor or to fine-tune it for some aligned applications, mainly to reduce training time and to encounter the overfitting problem while using a handful number of training samples. In this work, we use the Xception model [7] pre-trained on the ImageNet dataset as the backbone network following its success in other deepfake detection methods. Xception [7] is based on the widely used Inception architecture but with Inception modules replaced with depth-wise separable convolutions. In our implementation, we do not use the entire Xception model, rather we use the network consisting of a strided convolution block, followed by 12 depth-wise separable convolutions blocks with residual connections, except for the last one. Also, two depth-wise separable convolutions, a pooling operation and a batch normalization layer, are added to the network. The discussed model takes an input image of size (299×299) pixels and generates feature maps which are further passed to the later soft attention model (described in the following subsection) where useful features are selected for deep fake image classification.

4.3 Visual attention network

The ability of a neural network to learn discrete information elements to focus on within a given training sample was first proposed by Bahdanau et al. [36] in the field of machine translation. Since then, attention techniques has been useful in solving natural language processing (NLP) problems and computer vision problems. The benefit of using soft attention, in our work, is that it uses "soft shading" to focus on important regions of the image that would help in deciding fake from real faces. As a result, relevant features from the entire image are taken into consideration. On the other hand, a hard attention model uses image cropping to focus on the important regions in an image. As we use cropped faces in our work, manipulations in the entire image are of our interest and thereby using the soft attention model would be worthy. Soft attention is also end-to-end trainable with a gradient optimization method. Any soft attention mechanism primarily has two components - a network that learns probabilities for each information element within the input data, and a function that uses these probabilities to weight data for further processing.

The Xception model takes an input of dimension $(299 \times 299 \times 3)$ and generates features in the 3-D form $(W \times H \times D)$ where $W \times H$ and D denote the spatial resolution and the number of feature maps respectively. Let, there are $L = (W * H)$ number of locations, and the collection of these locations is represented by C which is represented by Eq. (2)

$$\mathbf{C} = \left\{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_L\right\}, \quad \mathbf{c}_i \in \mathbb{R}^D \tag{1}$$

In our work, the Xception model takes an input image of dimension $(299 \times 299 \times 3)$ and generates $(19 \times 19 \times 1024)$ dimensional features. Also, the dimension of C is (361×1024). Each element of C (i.e., context vector $(c_i, i = 1, 2, \ldots, L)$) is passed through the two branches. In one branch, linear transformation is applied on context vectors, while the other one keeps the copy of the original context vector. Let, after applying linear transformation (say, $lt()$) on c_i we obtain the transformed vector (say, t_i) i.e., $t_i = lt(c_i)$, $i = 1, 2, \ldots, L$. We call the collection of t_is as T which is a matrix of dimension $L \times D$. After this, we apply Eq. (2) on context vector C and transformed context T.

$$f_{\text{att}}(\mathbf{C}, \mathbf{T}) = \mathbf{M} * \tanh(\boldsymbol{U}\mathbf{T} + \boldsymbol{K}\mathbf{C} + \mathbf{b}) \tag{2}$$

In Eq. (2), M, U and K represent matrices of dimension $L \times D$, $D \times B$ and $D \times D$, respectively, while $b_{D\times 1}$ is a vector of dimension $D \times 1$. All these weight matrices and bias are learned during training process.

We get attention weights vector e_t by affine transformations followed by the logistic function on the context matrix C and the transformed matrix T as stated in Eq. (3).

$$e_t = f_{att}(\mathbf{C}, \mathbf{T}) \tag{3}$$

Equation (4) states α, the attention probabilities conditioned on the image context matrix C and transformed feature matrix T. The attention is then denoted as a vector of weights produced by a softmax function as shown in Eq. (4)

$$\alpha_{t,i} = \frac{\exp\left(e_{t,i}\right)}{\sum_{k=1}^{L} \exp\left(e_{t,k}\right)} \tag{4}$$

Getting the attention score vector $\alpha_{t,i}$, we terminate the model by stacking a couple of dense layers and a softmax layer for final classification. The attention weights and probabilities are generated with neural networks.

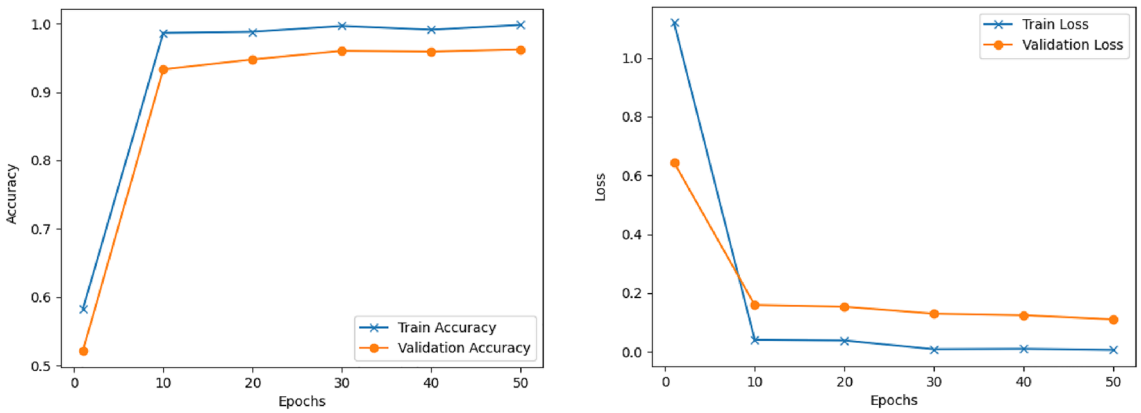
5 Results and discussion

We have evaluated the proposed model on two publicly available deepfake datasets, namely FaceForensics++ [11] and Celeb-DF (V2) [12]. The FaceForensics++ video dataset contains 1000 videos of each category-DeepFake, Face2Face, FaceSwap, NeuralTexture, and Pristine. So, the dataset contains 5000 videos in total. This dataset also provides 1000 unlabeled deepfake images for testing model performance on unseen datasets. Celeb-DF (V2) is a large-scale high-quality deepfake video dataset with 5639 synthetic videos of different celebrities. These videos are generated from 590 original videos (i.e., real videos) taken from YouTube with different subjects like genders, ages, and backgrounds. Thus, this dataset contains 6,229 videos out of which 518 videos are predefined as test set (178 real and 340 fake). We have first discussed different experiments to come up with an optimal trained network of our model, and then using this trained model, we have evaluated our model on unseen data and compared the results with state-of-the-art methods.

We train the entire model with a batch size of 32, and the step size per epoch is 40. We train the model for 50 epochs with a dynamic learning rate and Adam optimizer. The model is trained on the Nvidia Tesla K80 which consists of 4992 CUDA cores.

5.1 Experiments on FaceForensics++ dataset

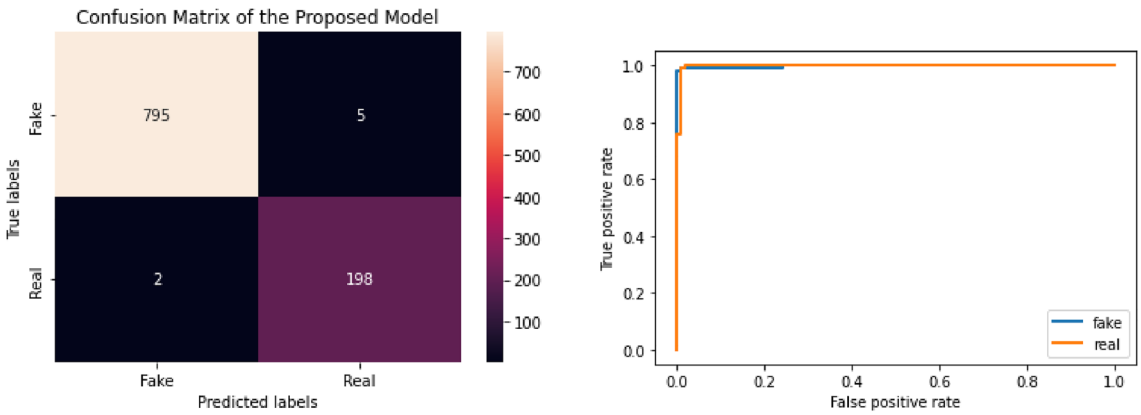
At first, we have partitioned the video samples of the FaceForensics++ dataset into two sets: train and test having 800 and 200 videos, respectively, from each category of deepfake videos. Next, we perform two sets of experiments to perform binary classification tasks, i.e., classifying fake and real videos. In the first set of experiments, we have considered all categories of manipulated videos (i.e., DeepFake, Face2Face, FaceSwap, and NeuralTexture) as fake videos and Pristine category videos as real videos. Thus, the



(a) Training and validation accuracy curve

(b) Training and validation loss curve

Fig. 4 Plots depicting training nature of the proposed method on FaceForensics++ dataset



(a) Confusion matrix

(b) ROC-AUC curve

Fig. 5 Analysis of metrics generated by evaluating the proposed method on test samples of FaceForensics++ dataset

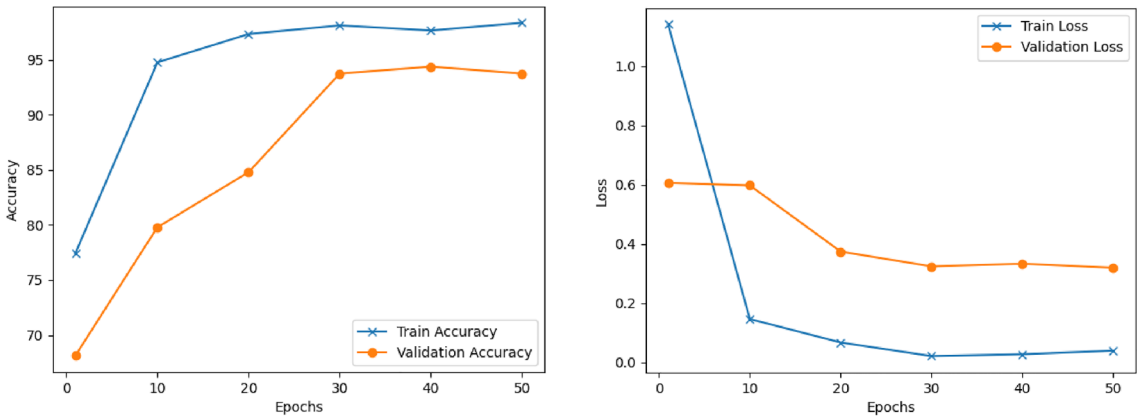
Table 1 Test accuracy of proposed model (with and without at tension network) on different categories of FaceForensics++ videos

Experiment	Test accuracy (in %)	
	Without attention	With attention
DeepFake vs Pristine	95.48	97.12
Face2Face vs Pristine	97.50	99.43
FaceSwap vs Pristine	96.00	96.45
NeuralTexture vs Pristine	96.12	98.00

In these experiments, videos of Pristine category are considered as real and the rest are fake

training set contains 3200 videos as fake and 800 videos as real. To tackle the class imbalance problem during training, we have extracted 4 and 1 P-frame(s) from each real and fake video, respectively, and thus, the training set consists of 3200 frames per class. However, we have extracted only one P-frame from each test video and thus in total 1000 frames used as test samples.

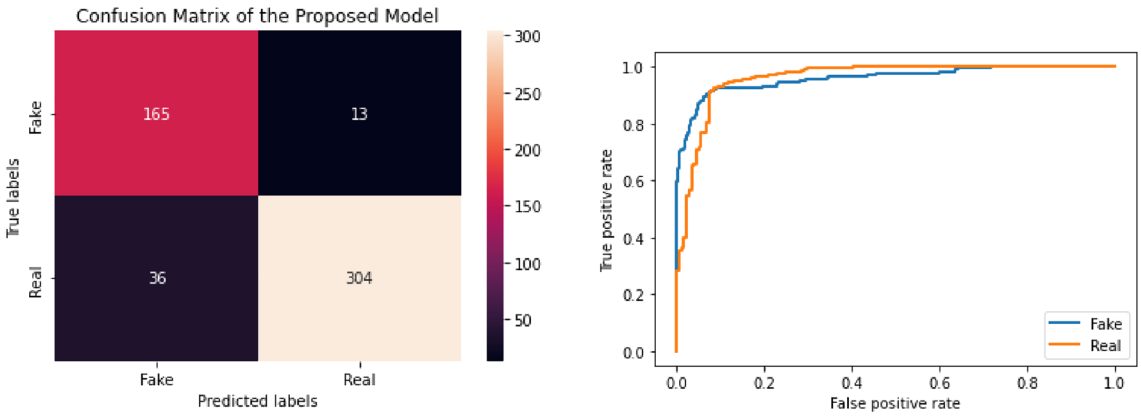
During model training, we have used 20% of the training samples as validation samples. The train and validation accuracies obtained with varying number of epochs are shown in Fig. 4a, while the train and validation loss scores with varying number of epochs are depicted in Fig. 4b. We obtain a proper nearly monotonous loss curve, which also indicates the model is trained properly. So, for the rest of the experiments, we train our model with 50 epochs. We have obtained 99.30% test accuracy while using proposed visual attention mechanism on the top of Xception network. While the test accuracy is decreased to 96.40% without using the proposed attention technique. Thus, use of the present attention mechanism helps to improve test accuracy by 2.90%.



(a) Training and validation accuracy curve

(b) Training and validation loss curve

Fig. 6 Plots depicting training nature of the proposed method on Celeb-DF (V2) dataset



(a) Confusion matrix

(b) ROC-AUC curve

Fig. 7 Analysis of metrics generated by evaluating the proposed method on Celeb-DF (V2) dataset

We also present the confusion matrix and ROC-AUC curve obtained by evaluating the proposed model on test set of FaceForensics++ dataset in Fig. 5.

In the second set of experiments, we have conducted four experiments: DeepFake vs Pristine, Face2Face vs Pristine, FaceSwap vs Pristine, NeuralTexture vs Pristine. In each experiment, videos belonging to the Pristine category are considered real, while the videos belonging to other categories are considered fake. For these experiments, we have extracted exactly one P-frame from each video, and thus, 1600 (800 per class) and 400 (200 per class) frames are considered in train and test sets, respectively. Table 1 shows experimental results. It is observed that the present model relatively struggles while identifying FaceSwap videos from Pristine category videos. However, from the experiments, it is clear that the use of the proposed visual attention network is proven to be beneficial.

5.2 Experiments on Celeb-DF (V2) dataset

In the Celeb-DF (V2) video dataset, we have extracted the first I-frame from 5639 synthetic (i.e., fake) videos and eight intermediate I-frames including the first one from 590 original videos to handle the class imbalance problem. Thus, we have a total of 10359 cropped face images (5639 fake and 4720 real). Here, 20% of these cropped faces are randomly selected as validation samples during model training. To prepare the test samples, we have extracted only the first I-frame from each video of test set. Thus, the test set contains only 518 frames, i.e., cropped face images. In Fig. 6, we present the accuracy and loss curves of the proposed

Table 2 Performance comparison of different models tested on the FaceForensics++ image benchmark platform

Method	DF	F2F	FS	NT	Pristine	Total
Fridrich et al. [37]	73.6	73.7	68.9	63.3	34.0	51.8
Cozzolino et al. [38]	85.4	67.8	73.7	**78.0**	34.4	55.2
Rahmouni et al. [39]	85.4	64.2	56.3	60.0	50.0	58.1
Bayar and Stamm [25]	84.5	73.7	**82.5**	70.6	46.2	61.6
Rossler et al. [21]	74.5	**75.9**	70.9	73.3	51	62.4
Back et al. [40]	71.8	68.6	63.1	70.7	56.2	62.5
Afchar et al. [20]	**87.2**	56.2	61.1	40.6	**72.6**	66.0
Proposed	78.2	72.3	52.4	74.7	70.0	**70.1**

In this table, DF, F2F, FS and NT represent DeepFake, Face2Face, FaceSwap and NeuralTexture, respectively

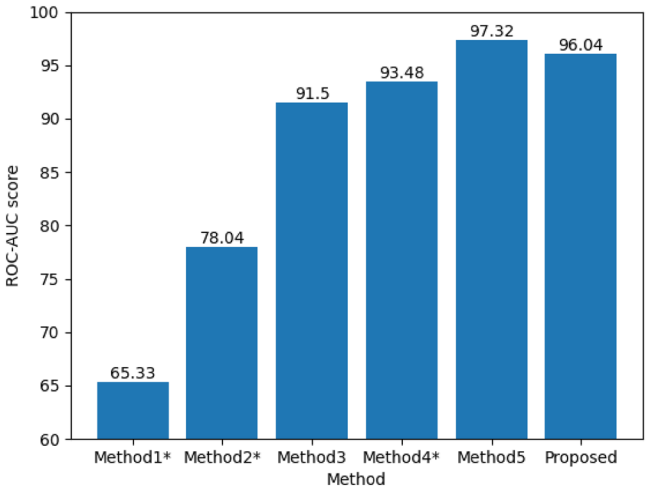


Fig. 8 Performance comparison of state-of-the-art methods with the proposed one using ROC-AUC score. Methods 1 to 5 denote the models proposed by Afchar et al. [20], Guo et al. [41], Ciftci et al. [42], Chollet et al. [7], Khormali and Yuan [43], respectively. * denotes trained and tested on our setup

method, depicting the variation of training and validation accuracies with varying number of epochs. In this case, like the previous dataset, we obtain a nearly monotonous training curve which confirms proper training, while the number of epochs is set to 50. The ROC-AUC scores obtained with and without using the proposed attention mechanism are 96.04% and 93.48%, respectively, while evaluated on present test samples. The confusion matrix of proposed deepfake detection method on this dataset is shown in Fig. 7a. We also present the corresponding ROC-AUC curve in Fig. 7b.

5.3 Comparison with state-of-the-art

To test the performance of the proposed method with state-of-the-art deepfake detection techniques, we have considered here some contemporary work [7, 20, 21, 25, 37–43]. The methods proposed by Fridrich et al. [37], Cozzolino et al. [38], Rahmouni et al. [39], Bayar and Stamm [25], Rossler et al. [21], Back et al. [40], and Afchar et al. [20] have evaluated performance of their models on the FaceForensics benchmark dataset that provides 1000 unlabeled target images to compare with state-of-the-art methods. In the benchmark images, the results are computed on a private server[1] by uploading predictions in a formatted JSON file. Hence, we have also followed the same approach while comparing performance of our model with these methods. In our case, we have used the trained model generated during the first set of experiments discussed earlier in Subsect. 5.1 to obtain the class information about the benchmark images. The comparative results are recorded in Table 2. We have obtained a total accuracy of 70.10% on the benchmark image dataset which is the best score among all the methods considered here for comparison. It is noteworthy to mention that following the state-of-the-art methods, we have considered the total score over other evaluation metrics mentioned in Table 2 for performance comparison as total score is the main trade-off among all mentioned scores in this table.

For performance comparison with existing methods on Celeb-DF (V2) dataset, we have compared our model performance with the methods proposed by Afchar et al. [20], Guo et al. [41], Ciftci et al. [42], Chollet et al. [7], and Khormali and Yuan [43]. Out of these methods, Ciftci

[1] http://kaldir.vc.in.tum.de/faceforensics_benchmark.

(a) Fake faces detected as Fake **(b)** Real faces detected as Real

(c) Fake faces detected as Real **(d)** Real faces detected as Fake

Fig. 9 Sample outputs obtained while evaluating the face images of the test set videos of FaceForensics++ dataset using proposed model. Here, we show **a** true positive, **b** true negative, **c** false positive and **d** false negative cases

et al. [42], and Khormali and Yuan [43] have evaluated their model performance on the test video set of Celeb-DF (V2) dataset. Thus, we have used the performances mentioned by these two methods for comparison. However, for the other methods, the same is not true, and hence, we have evaluated their performance using the current experimental set-up for fair comparison. Here, we have used the ROC-AUC score metric for performance comparison among the methods as suggested by Li et al. [12]. Fig. 8 shows the comparison of our model performance with the state-of-the-art methods on this dataset. On test data, we have obtained a 96.04% ROC-AUC score which outperforms the state-of-the-art methods except the method proposed by Khormali and Yuan [43]. However, this method was trained using over 2 million frames on Celeb-DF (V2) dataset, while our method uses only 8287 frames during its training process. Thus, it can be safely stated that the performance of the proposed model is at par with the state-of-the-art methods while comparing on Celeb-DF (V2) dataset.

5.4 Error analysis

In this section, we discuss and analyze the performance of the proposed model at the image level. For this analysis, we have considered the FaceForensics++ dataset. As discussed earlier, our model attention mechanism boosts the deep learning-based model to identify the artifacts left by the synthetic forged video generation process. After testing and visualizing the outcomes of the proposed model, we developed an intuition about the certain image features where the model may fail to act properly. We have discussed such observations here. Some of the correctly and erroneously classified images (i.e., extracted frames from videos) are shown in Fig. 9.

As shown in Fig. 9a, the model correctly classifies fake images where some artifacts are left by the system while generating deepfake images/videos. Also, from Fig. 9b, it is clear that the model correctly predicts real images which contain no artifacts or facial inconsistencies. However, the

model mostly fails to classify fake images which are generated by expression transfer methods like Face2Face, and NeuralTexture (see Fig. 9c). We have also observed that the model fails to classify some real images (see Fig. 9d) that have common features like closed eyes and some facial features, which the model considers as inconsistencies. We would also like to mention that we have evaluated our model on the C23 compressed version of FaceForensics++, which is moderately compressed, and thus, there is considerable information loss in the artifacts we look here for. Additionally, due to this heavy compression, some artifacts are redundantly generated which results in some misclassifications.

6 Conclusion

In this work, we have proposed a deep learning-based approach to detect manipulated videos that are generated with the help of some computer graphics and computer vision techniques. Here, we have applied a soft attention mechanism to learn specific synthetic artifacts in the fake videos that are absent in real ones. To the best of our knowledge, the proposed algorithm is the first founding step in the detection of video forgery where such an attention mechanism is applied. We have evaluated our model on two popularly used public deepfake-based forgery datasets, namely FaceForensics++ and Celeb-DF (V2), and our method outperforms many state-of-the-art methods used here for comparison.

Though the present model's performance is better than many state-of-the-art methods yet there is still some room for improvement. On the FaceForensics++ dataset, our method fails to provide competent results for FaceSwap and DeepFake types video manipulation in comparison with other two categories. In the future, more sophisticated attention mechanisms need to be tried to handle such issues. Here, we have used the Xception model as the backbone CNN model. We can try other CNN models along with the proposed visual attention technique. Besides, the proposed method can be applied on other datasets to test its robustness.

Acknowledgements We are thankful to the Center for Microprocessor Applications for Training Education and Research (CMATER) research laboratory of the Computer Science and Engineering Department, Jadavpur University, Kolkata, India for providing infrastructural support.

Author Contributions SG and SKM contributed to conceptualization, methodology, formal analysis and investigation, writing - original draft preparation, writing - review and editing, and resources; SM and RS were involved in conceptualization, methodology, formal analysis and investigation, writing - original draft preparation, writing - review and editing, resources, and supervision; EC contributed to conceptualization, writing - original draft preparation, writing - review and editing, and resources.

Funding Not Applicable in our case.

Availability of data and materials Not Applicable in our case.

Declarations

Conflict of interest/Competing interests The authors declare that they have no conflict of interest.

Ethics approval Not Applicable in our case.

Consent to participate Not Applicable in our case.

Consent for publication Not Applicable in our case.

Code availability Not Applicable in our case.

References

1. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. Adv Neural Inf Process Syst **27**
2. 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), pp 1–7. IEEE
3. Jung T, Kim S, Kim K (2020) Deepvision: deepfakes detection using human eye blinking pattern. IEEE Access 8:83144–83154
4. Tolosana R, Romero-Tapiador S, Fierrez J, Vera-Rodriguez R (2021) Deepfakes evolution: analysis of facial regions and fake detection performance. In: international conference on pattern recognition, pp 442–456. Springer
5. 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. IEEE
6. Walia S, Kumar K, Kumar M, Gao X-Z (2021) Fusion of handcrafted and deep features for forgery detection in digital images. IEEE Access 9:99742–99755
7. Chollet F (2017) Xception: deep learning with depthwise separable convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1251–1258
8. Nguyen HH, Yamagishi J, Echizen I (2019) Use of a capsule network to detect fake images and videos. arXiv preprint arXiv:1910.12467
9. Li Y, Yang X, Sun P, Qi H, Lyu S (2020) Celeb-df: a large-scale challenging dataset for deepfake forensics. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)
10. Pan D, Sun L, Wang R, Zhang X, Sinnott RO (2020) Deepfake detection through deep learning. In: 2020 IEEE/ACM international conference on big data computing, applications and technologies (bdccBDCAT), pp 134–143. https://doi.org/10.1109/BDCAT50828.2020.00001
11. Rossler A, Cozzolino D, Verdoliva L, Riess C, Thies J, Nießner M (2019) FaceForensics++: Learning to detect manipulated facial images. In: International conference on computer vision (ICCV)
12. Li Y, Yang X, Sun P, Qi H, Lyu S (2020) Celeb-df: A large-scale challenging dataset for deepfake forensics. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 3207–3216

13. Durall R, Keuper M, Pfreundt F-J, Keuper J (2019) Unmasking deepfakes with simple features. arXiv preprint arXiv:1911.00686
14. Guarnera L, Giudice O, Battiato S (2020) Deepfake detection by analyzing convolutional traces. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, pp 666–667
15. Moon TK (1996) The expectation-maximization algorithm. IEEE Signal Process Mag 13(6):47–60
16. Yang X, Li Y, Lyu S (2019) Exposing deep fakes using inconsistent head poses. In: ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 8261–8265. IEEE
17. Li H, Li B, Tan S, Huang J (2020) Identification of deep network generated images using disparities in color components. Signal Process 174:107616
18. Koopman M, Rodriguez AM, Geradts Z (2018) Detection of deepfake video manipulation. In: The 20th Irish machine vision and image processing conference (IMVIP), pp 133–136
19. Welch BL (1947) The generalization of 'student's'problem when several different population varlances are involved. Biometrika 34(1–2):28–35
20. Afchar D, Nozick V, Yamagishi J, Echizen I (2018) Mesonet: a compact facial video forgery detection network. In: 2018 IEEE international workshop on information forensics and security (WIFS), pp 1–7. IEEE
21. Rossler A, Cozzolino D, Verdoliva L, Riess C, Thies J, Nießner M (2019) Faceforensics++: learning to detect manipulated facial images. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 1–11
22. Shang Z, Xie H, Zha Z, Yu L, Li Y, Zhang Y (2021) Prrnet: pixel-region relation network for face forgery detection. Pattern Recog 116:107950
23. Sun K, Zhao Y, Jiang B, Cheng T, Xiao B, Liu D, Mu Y, Wang X, Liu W, Wang J (2019) High-resolution representations for labeling pixels and regions. arXiv preprint arXiv:1904.04514
24. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1–9
25. Bayar B, Stamm MC (2016) A deep learning approach to universal image manipulation detection using a new convolutional layer. In: Proceedings of the 4th ACM workshop on information hiding and multimedia security, pp 5–10
26. Amerini I, Galteri L, Caldelli R, Del Bimbo A (2019) Deepfake video detection through optical flow based cnn. In: Proceedings of the IEEE/CVF international conference on computer vision workshops, pp 0–0
27. Sun D, Yang X, Liu M-Y, Kautz J (2018) Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8934–8943
28. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556
29. Li Z, Sun Y, Zhang L, Tang J (2021) Ctnet: context-based tandem network for semantic segmentation. IEEE Trans Pattern Anal Mach Intell
30. Sun Y, Li Z (2021) Ssa: Semantic structure aware inference for weakly pixel-wise dense predictions without cost. arXiv preprint arXiv:2111.03392
31. Su Y, Xia H, Liang Q, Nie W (2021) Exposing deepfake videos using attention based convolutional lstm network. Neural Process Lett 53(6):4159–4175
32. Khormali A, Yuan J-S (2021) Add: attention-based deepfake detection approach. Big Data Cogn Comput 5(4):49
33. Wang C, Deng W (2021) Representative forgery mining for fake face detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 14923–14932
34. Chen S, Yao T, Chen Y, Ding S, Li J, Ji R (2021) Local relation learning for face forgery detection. Proc AAAI Conf Artif Intell 35:1081–1088
35. Jiang B, Ren Q, Dai F, Xiong J, Yang J, Gui G (2018) Multi-task cascaded convolutional neural networks for real-time dynamic face recognition method. In: International conference in communications, signal processing, and systems, pp 59–66. Springer
36. Bahdanau D, Cho K, Bengio Y (2014) Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473
37. Fridrich J, Kodovsky J (2012) Rich models for steganalysis of digital images. IEEE Trans Inf Forensics Sec 7(3):868–882. https://doi.org/10.1109/TIFS.2012.2190402
38. Cozzolino D, Poggi G, Verdoliva L (2017) Recasting residual-based local descriptors as convolutional neural networks: an application to image forgery detection. In: Proceedings of the 5th ACM workshop on information hiding and multimedia security, pp 159–164
39. Rahmouni N, Nozick V, Yamagishi J, Echizen I (2017) Distinguishing computer graphics from natural images using convolution neural networks. In: 2017 IEEE workshop on information forensics and security (WIFS), pp 1–6. https://doi.org/10.1109/WIFS.2017.8267647
40. Baek J-Y, Yoo Y-S, Bae S-H (2020) Generative adversarial ensemble learning for face forensics. IEEE Access 8:45421–45431. https://doi.org/10.1109/ACCESS.2020.2968612
41. Guo Z, Yang G, Chen J, Sun X (2021) Fake face detection via adaptive manipulation traces extraction network. Comput Vision Image Underst 204:103170. https://doi.org/10.1016/j.cviu.2021.103170
42. Ciftci UA, Demir I, Yin L (2020) Fakecatcher: detection of synthetic portrait videos using biological signals. IEEE Trans Pattern Anal Mach Intell. https://doi.org/10.1109/TPAMI.2020.3009287
43. Khormali A, Yuan J-S (2021) Add: attention-based deepfake detection approach. Big Data Cogn Comput. https://doi.org/10.3390/bdcc5040049