Multimedia Tools and Applications (2021) 80:18461–18478
https://doi.org/10.1007/s11042-020-10420-8

Detecting DeepFake, FaceSwap and Face2Face facial forgeries using frequency CNN

Aditi Kohli[1] · **Abhinav Gupta**[1]

Received: 14 May 2020 / Revised: 17 November 2020 / Accepted: 22 December 2020 /
Published online: 18 February 2021
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC part of Springer Nature 2021

Abstract
The face of a person plays a vital role in any communication or visual content. To enhance this visual content, popular and easy accessible editing tools are used. However, there malicious usage is spreading disharmony in the society, by tampering video evidences, defaming a person's image etc. Therefore a robust detection method is required to authenticate the visual content. Thus, a novel method is proposed to detect facial forgeries. The proposed method extracts faces from a target video and convert them into frequency domain using two dimensional global discrete Cosine transform (2D- GDCT). Thereafter, a 3 layered frequency convolutional neural network (fCNN) is employed to detect forged facial image. The proposed method is trained and tested on FaceForensics++ dataset and Celeb-DF(v2) dataset. In addition, its robustness is evaluated on standardized benchmark dataset and compared with the state-of-the-art methods to prove its effectiveness.

Keywords First keyword · Second keyword · More

1 Introduction

Recent advancements in neural networks have made generative adversarial network (GAN) popular among researchers. GAN is a powerful generator discriminator model, which learns from huge dataset and creates a new sample. This peculiar property of GAN has leveraged by many researchers to create versatile applications. GANs are used in generating new patterns in fashion industry [30], composing music [28], detecting anomaly in tumor images [23], constructing 3D objects from images [27], generating videos [13] and many more. The famous example in the context of videos is Talking Mona Lisa, a living portrait application developed by Samsung AI lab. Here, the video is generated using a single image of Mona Lisa. These development at one end is appreciating the research community, while on the other hand is raising concern in the society from its malicious usage. The consequences of the malicious usage get intensified, if it is done on the most vital part of visual content

✉ Abhinav Gupta
abhinav.gupta@jiit.ac.in

[1] Department of Electronics and Communication Engineering, Jaypee Institute of Information Technology, A-10, Sector-62, 201309, Noida, Uttar Pradesh, India

[9], i.e. a face of a person. The effects can be observed in the form of fake news, forged video evidences, and defaming a person. The popular example of such a malicious usage is DeepFakes [6], where GAN is used to denigrate a person by replacing the face with another person's face.

Facial manipulation techniques are either classified as expression based or identity based technique. In expression based technique, the facial expression of a person is transferred to another person. The most popular is Theis et al. [25] Face2Face, which is a real-time facial reenactment system. In Face2Face, facial expressions of real-time video is altered by reconstructing 3D face model. However, [24] is an offline expression manipulation technique, where a fake video is generated using audio input sequence. The [24] method learn lip movements from audio input sequence and map the expressions on source face.

In identity based techniques, the face (source) of a person is replaced by another face (target). The identity based manipulation is either a computer graphic based technique or GAN based technique. The most popular identity based techniques are FaceSwap [7] and DeepFakes [6]. The FaceSwap is a computer graphic based technique while DeepFake is a deep learning based technique. In FaceSwap, using computer graphics facial landmarks are detected and a 3D template is generated. This template is blended on a target face and the identity of a person is replaced. In DeepFake, is generated by rigorously training on the large dataset of source and target faces. It consists of two autoencoders with a shared encoder trained to map a source face on a target face. DeepFakes are widely used for face replacement forgery. Its public implementations are available, namely *FakeApp* and *FaceSwap github*.

The facial manipulation techniques like DeepFake [6], Face2Face [25], and FaceSwap [7] are so efficient that even human can get confused by these synthetic facial images. In [22], authors have conducted an experiment to evaluate the performance of humans to detect facial forgery. And it is observed that humans are unable to detect facial manipulations especially in practical scenarios (low quality videos or images). The above stated fact leads to spread distrust in the digital media and motivated researchers to work aggressively in the field of facial manipulation detection.

Researchers have used both machine learning and deep learning approaches to detect facial forgeries. In [18], Matern et al. have investigated the visual artefacts and derived handcrafted feature sets to detect generated faces, Deepfakes and Face2Face images. These feature sets are based on features from eyes, nose tip, teeth and face borders. Similarly, in [29] authors developed a handcrafted feature set considering inconsistency in head pose of Deepfake faces. However, with the advances in manipulation techniques, the visual artefacts observed in [18] and [29] become weaker and require deep learning based robust detectors.

Further, convolutional neural network (CNN) based features extractors are employed by authors to detect facial forgery. In [10], recurrent neural network (RNN) to detect DeepFake videos using CNN based features. Similarly, in [19] and [12] authors used different training strategy in CNN to detect the forged faces, like high pass filtering and transfer learning on facial forged dataset respectively. Zoung et al. [32] employed two stream neural networks for forged face detection, where one stream is a CNN based face classification stream and the other is a steganalysis based triplet patch stream. The target face is labeled forged after combining the scores from the two streams. The authors in [14] trained CNN to detect lack of eye blinking. Similarly, in [1] authors used CNN to detect missing details in eyes from an image. In [21], authors employed global pooling layers to extract spatial features of an image to detect forgery, while in [16] CNN detects the facial wrapping artefact incorporated due to manipulation techniques. Researchers have also employed complex neural architectures to detect these facial manipulation techniques. In [20], the

authors deployed capsule network to detect various facial forgeries. In [2], the authors have employed a generative adversarial ensemble learning method, where they train the generator and discriminator multiple times and focuses on improving the discriminator accuracy. Thus require rigorous training scenarios to achieve facial forgery detection. Another GAN based detector is employed by authors in [31], where a GAN simulator is designed to detect the artefacts introduced during generation of fake videos and images.

Thus, the researchers have observed that any facial manipulation technique leaves fingerprints which are detected either by machine learning or deep learning approach. Therefore, we propose a novel method to detect facial forgery by deploying frequency CNN. In the proposed method, the target video is first split into frames and further converted into frequency domain using 2D-GDCT. Thereafter, a convolutional neural network is trained to learn frequency features of a facial image. The proposed method is trained and evaluated on a FaceForensic++ dataset [22]. FaceForensics++ dataset is standard facial forgery detection dataset, which consists of videos from DeepFake [6], Face2Face [25], FaceSwap [7] and other manipulation techniques. The proposed fCNN has a simple 3 convolutional layered architecture as compared to GAN based complex architectures in [2] and [31]. Another advantage of proposed fCNN is that it does not require rigourous training scenarios as compared to [2]. Thus, the proposed fCNN is a simple and easily trainable detector for facial forgeries.

The key contribution of the paper is designing a novel frequency based CNN to classify between forged and pristine faces in DeepFake, FaceSwap and Face2Face videos. Further the activation maps of proposed fCNN are investigated to understand the key features responsible for the classification of faces. The proposed method is generalized on the most recent and challenging, Celeb-DF(v2) DeepFake dataset [17]. The proposed method is also compared on FaceForensics benchmark dataset with state-of-the-art methods to prove its efficacy.

This paper is organized as follows, Section 1 describes the introduction to facial forgery and related work. Section 2 emphasizes on the motivation behind the proposed method, followed by Section 3, which discusses details of proposed method. Sections 4 and 5 are related to dataset generation, experimentation, comparison, results and discussions. Finally, the conclusion of the paper is described in Section 6.

2 Motivation

A facial image has extensive information about structure and expression. Therefore, analyzing a facial image requires precise examination and in case of authentication, it becomes difficult. In the literature, face authentication is typically performed using spatial domain features. The authors use various spatial domain features of the face such as eyes, nose, mouth etc. to differentiate between forged and pristine face. However, frequency domain features are equally important. In [31] the authors have shown that spectrum based classifier are performing better than pixel based classifiers in detecting the forged images. The authors have designed GANs to synthesis artefacts and further added an up sampling component to their designed GANs to detect those artefacts. Application of frequency domain features is also seen in other authenticating scenario such as smartphones user detection. In [11] authors have designed two-stream network, which exploits frequency domain features to authenticate the legitimate user or an impostor on smartphones. Similarly, in [15] authors designed a frequency based method to authenticate the security of smartphones. Thus, inspired by the above mentioned successful use of frequency domain features, we propose

a frequency CNN (fCNN) to detect forged and pristine faces in DeepFake, FaceSwap and Face2Face videos.

The Fig. 1 explains how frequency domain features are altered in faces from pristine, DeepFake, FaceSwap and Face2Face videos of FaceForensics++ dataset [22]. The Fig. 1a-d present faces from pristine, DeepFake, FaceSwap and Face2Face videos respectively. The pristine and forged faces are converted into frequency domain by employing 2D global discrete cosine transform (2D-GDCT). The frequency domain representation of all the faces are visualized in colormap form in Fig. 1e-h. The dark red color shows high values while dark blue color shows relatively low values of 2D GDCT coefficients in colormap. The upper left corner presents low frequency sub bands and lower right corner presents high frequency sub-bands. Therefore, it can be visualized that magnitude of frequency sub bands are increasing across the main diagonal of 2D GDCT colormap.

On comparing pristine face (Fig. 1a) with DeepFake face (Fig. 1b) and FaceSwap face (Fig. 1c), it is visualized that face details are not clear and overlapping effects are present near eyebrows and cheeks regions. These artefacts are also observed in their corresponding colormaps. The value of 2D-GDCT coefficients in low frequency sub bands (i.e. the red color) are different in Fig. 1f and g as compared to their pristine counterparts in Fig. 1e. Similarly, high frequency sub bands coefficients are also altered. However in a Face2Face (Fig. 1d), only facial expressions are manipulated. Hence artefacts are seen near nose and mouth regions. These artefacts are also seen as alterations maximally occurring in low and middle frequency sub bands coefficient values in Fig. 1h. The above observations have motivated us to employ frequency based CNN for classifying between forged and pristine face.

3 Proposed method

The proposed method is divided into pre-processing and detection stage as shown in Fig. 2. The pre-processing stage consists of face extraction and frequency domain conversion.

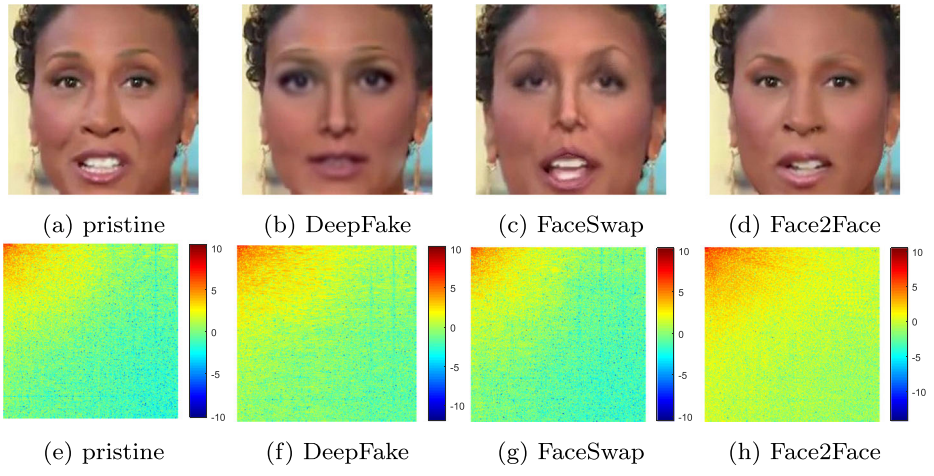

Fig. 1 **a-d** represent the faces from pristine (067.mp4), DeepFake (067_025.mp4), FaceSwap (067_025.mp4) and Face2Face (067_025.mp4) respectively of FaceForensics++ dataset [22]. **e-h** represent their corresponding 2D-GDCT coefficients in colormap

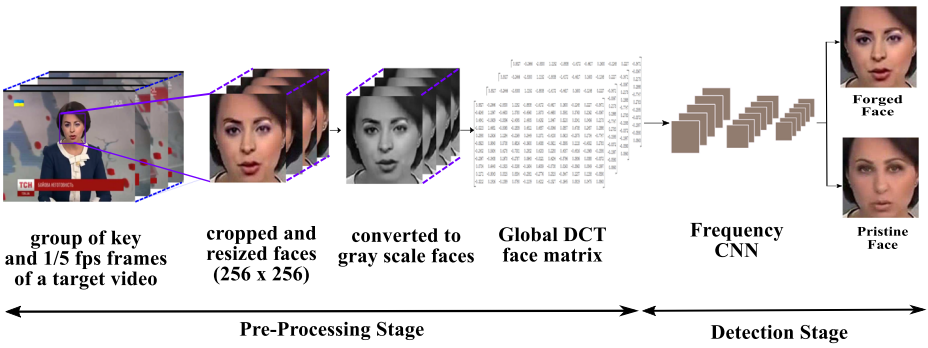

Fig. 2 Block diagram of proposed method

Detection stage consists of a 3 layered frequency convolutional neural network to classify between a forged and a pristine face.

3.1 Pre-processing stage

In this stage, the faces are extracted from target video and further converted into frequency domain using 2D global discrete cosine transform (2D-GDCT). The target video is split into frames. To avoid data redundancy, frames at 1/5 fps along with key frames are selected from a target video for further processing. This resultant group of frames are used for face extraction process. The face extraction is done with the help of Viola-Jones detector [26]. The extracted faces are resized to 256 x 256 and converted to gray scale image. Finally using 2D-GDCT, the faces are converted into frequency domain for the further processing in next stage.

Consider an image of size 256 x 256, where 8x8 blocked 2D-DCT is applied. Here, the image is first divided into blocks of size 8 x8 and then 2D-DCT is applied for each block. While, in case of 2D-GDCT, the DCT is applied on a complete image, i.e. the complete image is considered as a single block and then 2D-DCT is applied on it. Thus, the resulting DCT coefficients are equal to the size of an image.

The GDCT coefficients of an image represent information content in frequency domain. Let F be the facial image of size M x N, such that

$$F = [f(m,n), 0 \le m \le M-1, 0 \le n \le N-1] \tag{1}$$

and the GDCT of F be,

$$F_{gdct}(x,y) = a(x)a(y)\sum_{m=0}^{M-1}\sum_{n=0}^{N-1} f(m,n)\frac{cos(\pi x(2m+1))}{2M}\frac{cos(\pi y(2n+1))}{2N} \tag{2}$$

where,

$$a(x) = \begin{cases} \frac{1}{\sqrt{M}} & x = 0 \\ \sqrt{\frac{2}{M}} & 1 \le x \le M-1 \end{cases} \tag{3}$$

$$a(y) = \begin{cases} \frac{1}{\sqrt{N}} & y = 0 \\ \sqrt{\frac{2}{N}} & 1 \le x \le N-1 \end{cases} \tag{4}$$

3.2 Detection stage

The detection stage of the proposed method employs convolution neural network (CNN) to detect the forged faces in a target video. The CNN is an effective tool for analyzing computer vision and pattern recognition problems. It learn spatial hierarchies of features from low to high level patterns. The CNN is comprised of convolutional layer, activation layer, an optional pooling layer followed by fully connected layer. The convolutional, activation and pooling layers are responsible for feature extraction, while fully connected layer maps extracted features into final output, i.e. classification. The proposed method utilizes 3 convolutional layers to train the network for detection of the forged faces. A 2D-GDCT matrix, extracted in the pre-processing stage, corresponding to an input face is fed as an input to the proposed CNN. Therefore, the proposed CNN is named as frequency CNN (fCNN).

The fCNN is a 3 layered convolutional neural network as shown in Fig. 3. The input stage is a two dimensional frequency matrix corresponding to the gray scaled facial image. The first convolutional layer has 16 kernels of size 5×5, which are sliding at a stride rate of 1 to extract low level frequency features of a face. The linear output of convolutional layer is passed to non-linear activation layer which consists of ReLU activation function ($F(x) = max(0, x)$). Further, pooling layer is used to reduce the dimensions of output matrix (feature maps). It also reduces number of trainable parameters and make network invariant to translational shifts and disturbances. In proposed fCNN, maximum pooling technique is used among the two popular pooling techniques based on experimentation (discussed in Section 5). The 2×2 maximum pooling technique is employed at a stride rate of 2. This combination of convolutional, ReLU and maximum pooling layer corresponds to first layer of fCNN. Similarly, second layer of fCNN comprises of convolutional kernel of size $3 \times 3 \times 32$ and a ReLU activation layer followed by 2×2 maximum pooling layer. Third layer has convolutional kernel of size $3 \times 3 \times 48$. These convolutional kernels extract low to high level frequency features from input 2D GDCT matrix. Therefore, kernels are named as frequency kernels. The fCNN has 96 (16 + 32 + 48) frequency kernels, that play an important role in classification between a forged face and a pristine face. Their role is further elaborated in Section 5.1 with the help of activation maps.

Finally, all the frequency features are flattened in an array of size 1024 with dropout probability of 0.5 to avoid over-fitting of the network. Thereafter, two class classification is performed using softmax function. Hence, decision probabilities of softmax function of fCNN will classify 2D-GDCT matrix input into either a forged face or a pristine face.

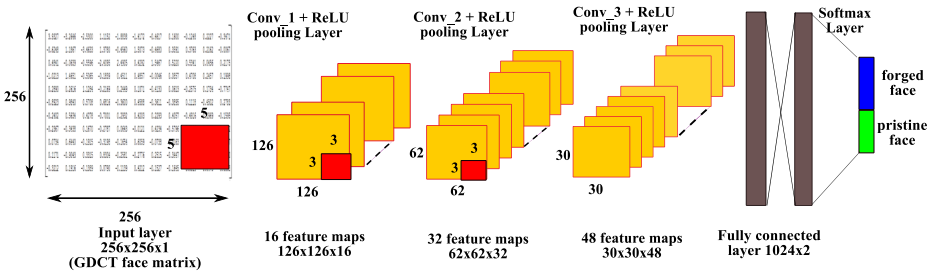

Fig. 3 The detailed architecture of fCNN. The red square represents the frequency kernels. The yellow squares at each layer of fCNN represent output after every layer (conv + ReLU + max pooling). The output of softmax layer is presented in blue and green color representing forged and pristine face respectively

The Fig. 4 presents the flow diagram of training and detection process of fCNN. The 2D-GDCT input sample (train) is passed through the feed forward network of fCNN. The error is back propagated through the network using stochastic gradient descent with momentum (SGDM) optimization algorithm. The SGDM is using momentum of 0.9 and learning rate of 0.001. The network is trained for 40 epochs. The training accuracy and loss with respect to number of epochs are presented in Fig. 5a and b respectively, for DeepFake, FaceSwap and Face2Face for $c23$ video quality. Appropriate training of the proposed fCNN network can be seen from Fig. 5.

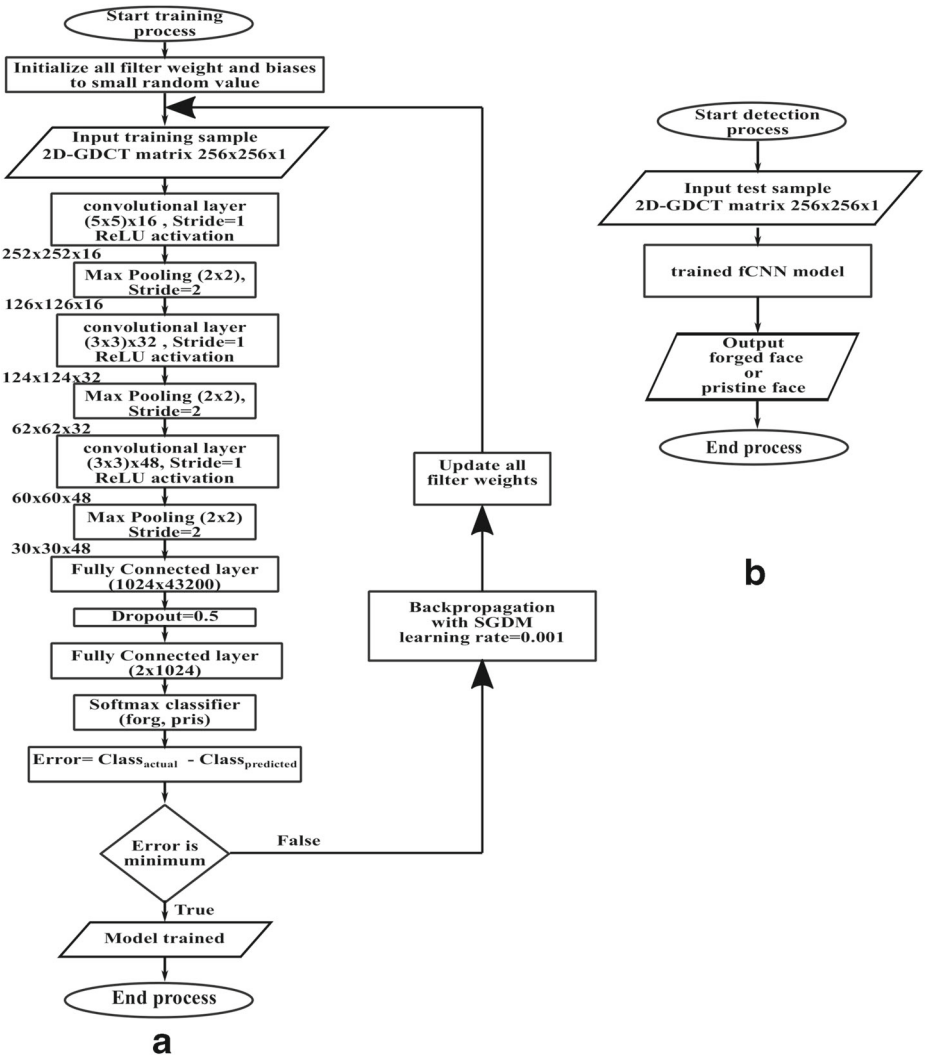

Fig. 4 **a** The flow diagram of proposed fCNN training process. **b** The flow diagram of proposed fCNN detection process

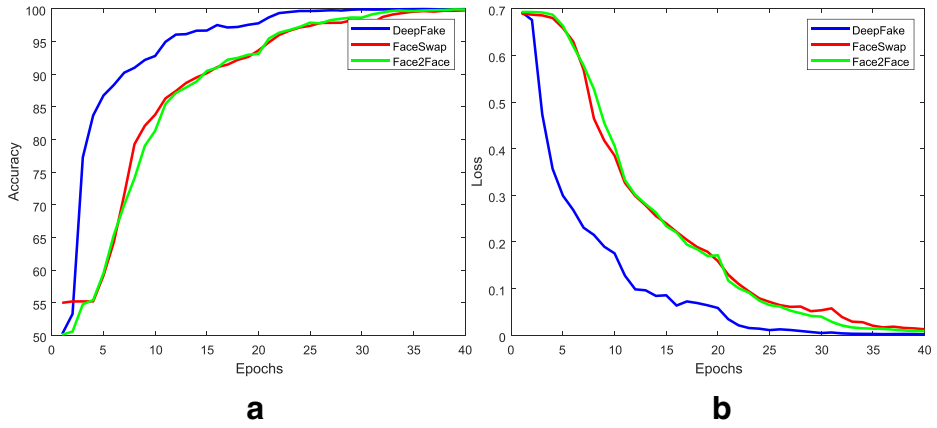

Fig. 5 **a** and **b** represent the training accuracy curve and loss curve respectively, of proposed fCNN over DeepFake $c23$, FaceSwap $c23$ and Face2Face $c23$ dataset

4 Dataset

The proposed method is trained and tested on FaceForensics++ dataset [22]. The FaceForensics++ dataset is a facial forgery dataset. It contains forged facial videos generated from four methods, namely, 'DeepFakes', 'FaceSwap', 'Face2Face' and 'Neural Textures'. It consists of 1000 pristine, 1000 DeepFake, 1000 FaceSwap, 1000 Face2Face and 1000 Neural Textures videos. The dataset has videos with resolution 480p, 720p and 1080p. These 5000 videos (pristine and forged) are compressed at different quality factors, namely, raw, high quality (HQ) and low quality (LQ). The raw videos are uncompressed, the HQ videos have low compression with constant quantization parameter 23 ($c23$), and the LQ videos have high compression with constant quantization parameter 40 ($c40$). Both male and female faces are included in the dataset. Therefore, FaceForensics++ dataset is a large scale generic facial dataset.

The proposed method is evaluated for three facial manipulations. Therefore, three datasets are derived from FaceForensics++ dataset [22], namely DeepFake dataset (consists of 1000 DeepFake and 1000 orginal videos), FaceSwap dataset (consists of 1000 FaceSwap and 1000 orginal videos) and Face2Face dataset (consists of 1000 Face2Face and 1000 orginal videos). All the datasets are divided into training, validation and testing groups, which consists of 700, 150, 150 videos respectively. The videos are split into frames using ffmpeg tool and thereafter the faces are extracted. The extracted faces are converted to gray scaled image and are re-sized to 256×256. The cardinality of each class in different datasets are discussed in Table 1.

The performance of proposed method is evaluated in terms of accuracy (Ac), precision (P), Recall(R) and $F1score$. Ac measures number of correctly detected faces from total number of faces, while P tells about exactness of forged detected faces (positive predicted value). R measures completeness of forged detected faces (sensitivity or true positive rate) and $F1score$ is harmonic mean of precision and recall. $F1score$ is considered best, when it is 1. The above stated metrics are calculated using following equations:

$$Ac = \frac{TP + TN}{TP + TN + FN + FP} * 100 \tag{5}$$

Table 1 The cardinality of each class in the three datasets, namely DeepFake, FaceSwap and Face2Face

	DeepFake		FaceSwap		Face2Face		Pristine	
Dataset	Training	Testing	Training	Testing	Training	Testing	Training	Testing
RAW	5386	819	4430	693	5430	820	5380	827
HQ	5362	817	4435	694	5461	822	5463	843
LQ	5379	824	4424	687	5443	825	5348	814

$$P = \frac{TP}{TP + FP} \tag{6}$$

$$R = \frac{TP}{TP + FN} \tag{7}$$

$$F1 = 2\frac{P \times R}{P + R} \tag{8}$$

where, TP is number of forged faces detected as forged, TN is the number of pristine faces detected as pristine, FN is the number of forged faces detected as pristine and FP is the number of pristine faces detected as forged. The AUC (Area under the ROC curve) measure is also included to measure the discriminability of forged and pristine faces.

5 Results and discussion

In this section, the network parameter selection and the robustness of proposed method is discussed in detail. The proposed fCNN is evaluated for different facial manipulation scenarios, i.e. DeepFake, FaceSwap and Face2Face. Further to understand how frequency based convolutional network is classifying the faces, activation maps are also studied in detail. The proposed method is evaluated on Celeb-DF(v2) dataset [17]. Finally, the proposed method is benchmarked on the publicly available facial manipulation detection dataset [22]. All the experiments are performed on Xeon W-2123 CPU, 3.60 GHz with 16GB RAM using MATLAB R2018b.

The proposed fCNN is designed to classify between DeepFake and pristine faces. The videos at HQ (c23) quality are quite similar to publicly available videos on social networks. Therefore, all the experiments are performed on DeepFake HQ (c23) quality videos. Intensive experiments are conducted to select the network parameters: number of filters, pooling techniques, number of trainable parameters, dropout probability and batch size. The Table 2 presents the validation and testing accuracy for three different sets of number of filters in convolutional layers, i.e. (8, 16, 24), (16, 32, 48), and (32, 48, 96). The listed results represent that (32, 48, 96) number of filters give approximately 1% better testing accuracy than

Table 2 The validation and testing accuracy for different number of filters of 3 layered frequency CNN

No. of filters	Validation *Ac*	Testing *Ac*	Trainable parameters
8,16,24	92.69	85.06	22126330
16,32,48	**93.99**	85.24	44258802
32,48,96	93.44	**86.08**	88532946

Best achieved accuracy are shown in bold

Table 3 The validation and testing accuracy for average and maximum pooling technique

Pooling technique	Validation *Ac*	Testing *Ac*
avg	87.88	84.76
max	**93.99**	**85.24**

Best achieved accuracy are shown in bold

(16, 32, 48) set. However, the trainable parameters are almost double in case of (32, 48, 96) set. Therefore, (16, 32, 48) number of filters are chosen for the 3 layered fCNN architecture.

The experiments are performed for the two popular pooling technique, i.e. average and maximum pooling. The results listed in Table 3 present that in case of maximum pooling both training and testing accuracy are better. The dropout is employed in any network to avoid over fitting of the network parameters. The experiments are also conducted for various dropout probabilities like 0.2, 0.3, 0.4, 0.5. The Table 4 results show improvement in the performance wih 0.5 dropout probability. The network is validated and tested for different batch sizes, i.e 28 and 32. The results listed in Table 5 clearly show the batch size 32 as a best candidate.

The computation complexity and computational cost of the proposed network is analyzed using model size and number of floating point operations (FLOPs) [4], respectively. The model size is measured using total number of trainable parameters in the proposed network. The details about total trainable parameters along with the output shape for each convolutional layer is listed in Table 6, as model summary. However, the number of FLOPs are calculated using number of multiply-adds (MACCs). The MACCs are computed for convolutional (conv) and fully connected (FC) layer using given equations,

$$MACCs(conv) = d_t^2 C_t d_k^2 C_s \tag{9}$$

$$MACCs(FC) = C_t C_s \tag{10}$$

where, d_k is kernel spatial size, d_t is output feature map size, C_t is number of output channels and C_s is number of input channels. Further, the calculated MACCs for each conv and FC layer are added to give total number of MACCs. The multiply-adds (MACCs) are counted as two FLOPs [4]. Thus, for proposed fCNN there are 44 M parameters and 380 M-FLOPs.

The proposed method is evaluated for detection of three facial manipulation techniques along with three compression qualities. To achieve this, the proposed fCNN is specifically trained for three types of DeepFake datasets, i.e. raw DeepFake, $c23$ DeepFake, $c40$ DeepFake. Similar training is also done for FaceSwap and Face2Face datasets. For different video qualities, the validation accuracy, testing accuracy and AUC of the three manipulation techniques are listed in Table 7. The compressed videos (c40) have less high frequency component, thus their detection accuracy is least for all the three manipulations as observed

Table 4 The validation and testing accuracy for different dropout *Pe*

Dropout *Pe*	Validation *Ac*	Testing *Ac*
0.2	93.88	85.02
0.3	92.69	85.03
0.4	91.77	84.64
0.5	**93.99**	**85.24**

Best achieved accuracy are shown in bold

Table 5 The validation and testing accuracy for different batch sizes

Batch size	Validation Ac	Testing Ac
28	92.88	85.54
32	**93.99**	**85.24**

Best achieved accuracy are shown in bold

in Table 7. For uncompressed data (raw) comparing DeepFake, FaceSwap and Face2Face detection accuracies, the testing accuracy of FaceSwap manipulation detection is higher than other two manipulation detections. However, in case of $c23$ and $c40$ video quality, the DeepFake testing accuracy and AUC are better. Further, in Table 8 precision (P), recall (R) and $F1score$ metrics are listed for all the above stated scenarios. The results show that the better precision (positive predicted value) and recall (true positive rate) of the proposed fCNN for detecting DeepFake manipulation as compared to FaceSwap and Face2Face manipulations. The harmonic mean of precision and recall, i.e. the $F1$ score is also good for DeepFake manipulations in case of $c23$ and c40 compressions. Hence it can be inferred that proposed fCNN is distinguishing the DeepFake facial forgery better.

The proposed fCNN is also trained and tested on Celeb-DF(v2) dataset [17], that is the recent and challenging DeepFake video dataset. It consists of 5639 DeepFake videos synthesized from publicly available YouTube video clips of 59 celebrities of diverse genders, ages, and ethic groups. The computed testing Ac, AUC, P, R and $F1score$ are tabulated in Table 9. The results show a good precision (0.9290) of proposed fCNN in detecting DeepFake faces of Celeb-DF(v2).

Table 6 fCNN model summary

Layer (type)	Output Shape	Trainable Parameters
Input Layer	256 x 256 x 1	0
conv2d_1 (Conv2D)	252 x 252 x 16	416
activation_1 (Activation)	252 x 252 x 16	0
max_pooling_2d_1	126 x 126 x 16	0
conv2d_2 (Conv2D)	124 x 124 x32	4640
activation_2 (Activation)	124 x 124 x32	0
max_pooling_2d_2	62 x 62 x 32	0
conv2d_3 (Conv2D)	60 x 60 x 48	13872
activation_3 (Activation)	60 x 60 x 48	0
max_pooling_2d_3	30 x 30 x 48	0
flatten_1	43200	0
dense_1	1024	4423782
dropout_1	1024	0
dense_2	2	2050
Total trainable parameters		**44,258,802**

Total number of parameters are shown in bold

Table 7 The binary validation accuracy(%) and testing accuracy(%) and AUC(%) of proposed fCNN for DeepFake, FaceSwap and Face2Face manipulation techniques on FaceForensic++ dataset [22]

	DeepFake			FaceSwap			Face2Face		
Dataset	Validation	Testing		Validation	Testing		Validation	Testing	
	Ac	*Ac*	AUC	*Ac*	*Ac*	AUC	*Ac*	*Ac*	AUC
raw	91.78	87.79	95.96	91.64	**89.28**	**96.70**	88.62	85.37	93.12
c23	93.99	**85.24**	**93.34**	86.54	85.03	93.10	88.28	81.08	89.06
c40	89.19	**79.24**	**88.41**	79.12	69.35	76.02	88.90	68.88	77.74

Best detection accuracy are shown in bold

5.1 Visualizing fCNN activation maps

The activation maps of trained CNN are helpful in understanding and visualizing the classification performance for a dataset. The proposed fCNN is therefore investigated in the context of activation maps. Fifty random DeepFake and pristine faces are selected from DeepFake c23 test dataset of FaceForensics++ dataset [22]. Then, their resultant activation maps of size $60 \times 60 \times 48$ are extracted from the last convolutional layer of fCNN. Thereafter, 4 filters are randomly selected and presented using colormaps in Fig. 6. In activation colormap, the dark red color presents strong activation whereas dark blue color presents weak activation. The Fig. 6a-d represents the DeepFake activation maps and Fig. 6e-h represents their corresponding pristine activation maps. In Fig. 6b the middle and high frequency sub bands are slightly activated while in Fig. 6f high frequency band is weakly activated. Similarly, for low frequency sub bands, Fig. 6c shows weakly activation for while its counterpart in Fig. 6g shows strong activation.

Similarly, activation maps are extracted for FaceSwap and Face2Face forgeries. A set of 4 filters are randomly selected and presented using colormap in Figs. 7 and 8 respectively. The strongly activated frequency bands in Fig. 7a - d are completely different from those activated in Fig. 7e - h. Furthermore, the activation maps of Face2Face forgery also presents the distinct frequency band activations compared to their corresponding pristine activation maps. In Fig. 8a colormap presents orange color for middle frequency band while its counterpart in Fig. 8e is presented by mixture of yellow and green color. Thus, it can be inferred that the proposed fCNN is learning different frequency sub bands to classify the forged and pristine faces.

Table 8 The precision P, recall R and $F1$ score for DeepFake, FaceSwap and Face2Face manipulation detection on FaceForensics++ datasest [22]

	DeepFake			FaceSwap			Face2Face		
Dataset	*P*	*R*	*F*1 score	*P*	*R*	*F*1 score	*P*	*R*	*F*1 score
raw	0.8205	**0.9256**	0.8699	**0.9148**	0.8591	**0.8861**	0.8756	0.8378	0.8563
c23	**0.8311**	**0.8639**	**0.8472**	0.8040	0.8598	0.8310	0.8029	0.8118	0.8073
c40	**0.7257**	**0.8399**	**0.7786**	0.7205	0.6488	0.6828	0.6630	0.7022	0.6820

Best P, R, F1 score for particular dataset are shown in bold

Table 9 The testing result of proposed fCNN on Celeb-DF(v2) dataset [17]

Dataset	Testing Ac	AUC	P	R	$F1$ score
Celeb-DF(v2)	66.50	75.20	0.9290	0.6172	0.7394

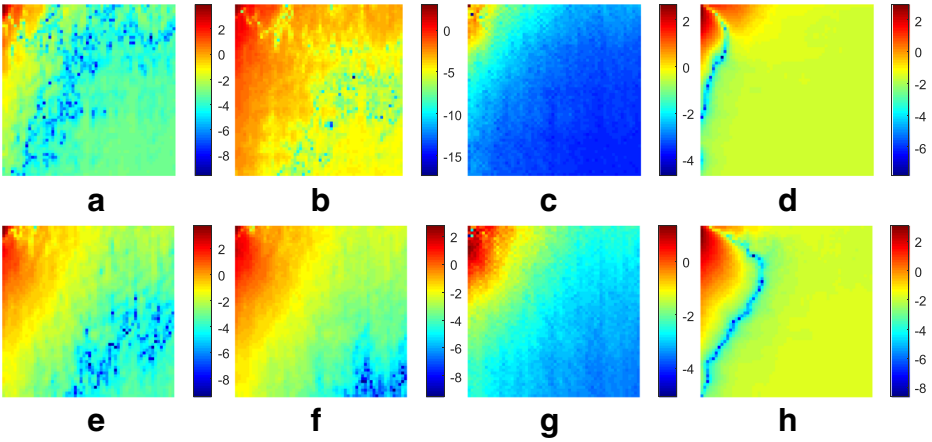

Fig. 6 Mean activation maps of some randomly selected filters of last convolutional layer of fCNN of size 60×60 for DeepFake $c23$ test dataset of FaceForensics++ dataset [22]. **a**-**d** represent DeepFake and **e**-**h** represent pristine activation maps

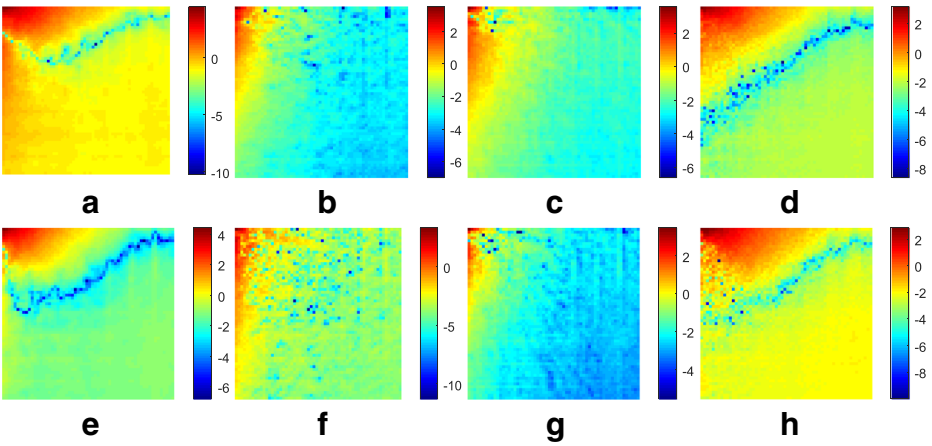

Fig. 7 Mean activation maps of some randomly selected filters of last convolutional layer of fCNN of size 60×60 for FaceSwap $c23$ test dataset of FaceForensics++ dataset [22]. **a**-b represent FaceSwap and **e**-**h** represent pristine activation maps

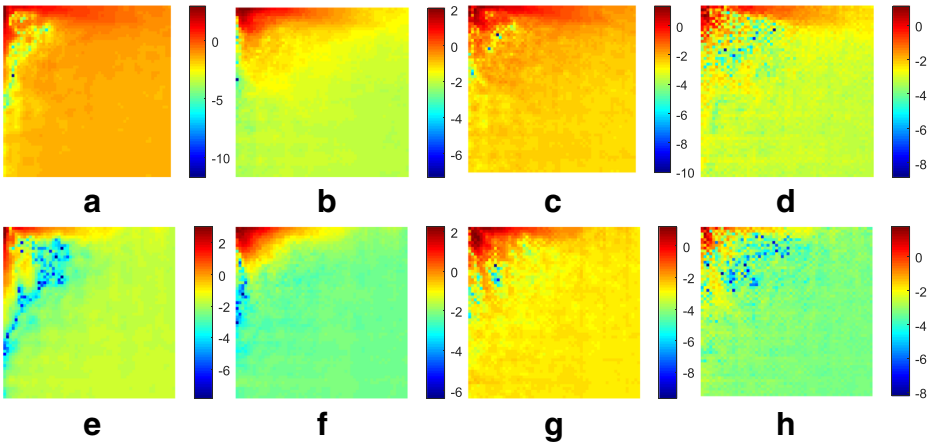

Fig. 8 Mean activation maps of some randomly selected filters of last convolutional layer of fCNN of size 60×60 for Face2Face $c23$ test dataset of FaceForensics++ dataset [22]. **a**-**d** represent Face2Face and **e**-**h** represent pristine activation maps

5.2 Detecting facial forgery technique from forged face: multi-class detection

In literature, researchers have paid scant attention to detection of facial forgery techniques. In [1, 21] and [14], authors worked to detect whether the face is forged or pristine. Therefore, to know about the facial manipulation technique incorporated by forger, we evaluated the proposed method for multi-class facial forgery detection. The proposed fCNN is trained and fine tuned for multiple classes, namely DeepFake, Face2Face, FaceSwap and pristine. The network is trained with 20641 facial images that comprises of 5362 DeepFake, 5461 Face2Face, 4435 FaceSawp and 5383 pristine images. The training dataset is split into training and validation group. The 10% of training dataset is employed for validating the network. The testing results in form of confusion matrix is discussed in Table 10. The proposed method is identifying the multiple facial forgeries with overall accuracy of 78.3%. The best detected facial manipulation technique is DeepFake followed by FaceSwap and Face2Face. Therefore, the proposed method is able to detect the forged face as well as give details about the forgery technique used by the forger to manipulate the face.

Table 10 Confusion matrix for detecting DeepFake, Face2Face, FaceSwap and Pristine facial manipulation techniques

		True Class			
		DeepFake	Face2Face	FaceSwap	Pristine
Predicted Class	DeepFake	88.86%	9.98%	1.44%	14.59%
	Face2Face	3.55%	73.36%	3.03%	9.25%
	FaceSwap	0.00%	3.41%	83.86%	7.83%
	Pristine	7.47%	13.26%	11.67%	68.33%

5.3 Comparison on faceforensics benchmark

In [22], authors have proposed an automated benchmark for facial manipulation detection. The benchmark dataset consists of 1000 facial images either taken from original or manipulated videos (namely DeepFakes, FaceSwap, Face2Face and Neural Textures), at different compressions (raw, $c23$, $c40$). The ground truth of benchmark dataset is unknown to submitter, to prevent over fitting of network. To test the facial detection method on the benchmark dataset, the predicted labels corresponding to the 1000 images in '.json' file format are uploaded on third party server. The server then evaluated the submitted label and displayed the detection accuracy on the benchmark website.

The robustness of proposed method is evaluated on above discussed FaceForensics++ benchmark dataset [22]. Therefore a mixed dataset is derived from FaceForensics++ dataset [22], to train the proposed fCNN for benchmark purpose. The mixed dataset consists of 16191 forged and 16200 pristine facial image classes. The pristine facial class comprises of 5380 raw, 5463 $c23$ and 5438 $c40$ images. While forged facial class comprises of 1350 randomly selected images each from DeepFake, Face2face, FaceSwap and Neural Textures manipulation techniques for all compression qualities, i.e. raw, $c23$, $c40$ (1350×12). The 10% of the above stated training dataset is employed for validating the proposed fCNN.

The Table 11 presents a comparison of proposed fCNN with recent state-of-the-art detectors. The XceptionNet [22] is performing best among detectors in terms of total detection accuracy, but the proposed fCNN outperforms the XceptionNet [22] in terms of detecting pristine faces. Similarly, the GAEL-Net [2] which is also better than proposed fCNN in total detection accuracy, shows a performance reduction when detecting DeepFake and FaceSwap forgeries as compared to proposed fCNN. Thus, upon inspection of detectable accuracies listed in Table 11, it can be said that no single detector is the best at detecting all five classes.

In addition, the detectors described in Table 11 are biased to a particular forgery. Therefore, as suggested in [2], the standard deviation score for all five detection accuracy is calculated and listed in Table 11, where a lower standard deviation means a less biased detector toward a specific forgery. It is observed from Table 11 that, GAEL Net [2] and XceptionNet Full image [22] have lower standard deviation score than proposed

Table 11 The comparison of state-of-art facial forgery detectors with proposed fCNN for DeepFake (DF), Face2Face (F2F), FaceSwap (FS) and Neural Textures (NT) and pristine (Pris) detection accuracy

	DF	F2F	FS	NT	Pris	Total	Std. Dev.
Steganalysis Features [8]	0.736	0.737	0.689	0.633	0.340	0.518	0.166
Recasting [5]	0.855	0.679	0.738	0.780	0.344	0.552	0.198
Rahmouni [21]	0.855	0.642	0.563	0.607	0.500	0.581	0.135
Bayar and Stamm [3]	0.845	0.737	0.825	0.707	0.462	0.616	0.153
XceptionNet Full Image [22]	0.745	0.759	0.709	0.733	0.510	0.624	0.103
MesoNet [1]	0.873	0.562	0.612	0.407	0.726	0.660	0.175
Xception [22]	0.964	0.869	0.903	0.807	0.524	0.710	0.171
GAEL Net [2]	0.718	0.686	0.631	0.707	0.562	0.625	0.065
fCNN (proposed method)	0.791	0.642	0.709	0.513	0.544	0.597	0.115

The standard deviation (Std. Dev.) scores are also compared for five accuracies

fCNN. However, it can be inferred from Table 11 that there is a trade-off between standard deviation and detection accuracies. Furthermore, comparing the proposed fCNN with XceptionNet [22] and GAEL Net [2] in terms of architecture and training scenarios gives another perspective. The XceptionNet is a 36 convolutional layered deep architecture while the proposed fCNN is a three convolutional layered shallow architecture. Similarly, GAEL Net [2] is a GAN based architecture that requires rigorous training than a simple CNN based architecture of the proposed fCNN. Therefore, it can be deduced that the proposed simple architecture fCNN is performing reasonably well in detecting most of the facial manipulation techniques.

The limitation of proposed method is its non-uniform facial manipulation detection accuracy. For detection of neural texture facial manipulation, the detection accuracy (0.5130) is approximately a random guess. Thus, the future work can be extended to improve the facial detection accuracy for neural textures, as well as to design an unbiased detector for all types of facial forgeries.

6 Conclusion

High quality face editing tools are spreading distrust in digital media, especially DeepFakes. Over the past two years, researchers worked extensively in spatial domain techniques to detect facial forgery. The proposed method is employed to exploit the frequency domain features of the facial forgery. The proposed method deploy a frequency CNN (fCNN) to analyze and classify between pristine and forged face. FaceForensics++ dataset is used to evaluate the effectiveness of the fCNN. Experiments show that fCNN is effectively detecting forgery in practical scenarios, i.e. HQ and LQ video qualities. In addition, visualization of activation maps shows that distinct frequency features are learned by fCNN for DeepFake, Face2Face and FaceSwap manipulation techniques.

The proposed fCNN is detecting DeepFakes with highest recall of 0.9256, 0.8639 and 0.8399 for raw, $c23$ and $c40$ respectively, among all other facial manipulation techniques. The proposed method is also evaluated on a Celeb-DF(v2) dataset and an automatic FaceForensic benchmark. The benchmark results of the proposed method are compared to state-of-the-art methods. The comparison results prove the efficacy of the proposed method for facial manipulation detection.

References

1. 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
2. Baek J, Yoo Y, Bae S (2020) Generative adversarial ensemble learning for face forensics. IEEE Access 8:45421–45431
3. Bayar B, Stamm M (2016) A deep learning approach to universal image manipulation detection using a new convolutional layer. pp 5–10. https://doi.org/10.1145/2909827.2930786
4. Bianco S, Cadene R, Celona L, Napoletano P (2018) Benchmark analysis of representative deep neural network architectures. IEEE Access 6:64270–64277
5. 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, IH&MMSec '17, pp 159–164, Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3082031.3083247

6. Deepfakes github. https://github.com/deepfakes/faceswap. Accessed: 2020-02-01
7. Faceswap github. https://github.com/MarekKowalski/FaceSwap. Accessed: 2020-02-01
8. Fridrich J, Kodovsky J (2012) Rich models for steganalysis of digital images. IEEE Trans Inform Forensics Secur 7(3):868–882. https://doi.org/10.1109/TIFS.2012.2190402
9. Frith C. (2009) Role of facial expressions in social interactions. philosophical transactions of the royal society of london. Series B Biol Sci 364:3453–8. https://doi.org/10.1098/rstb.2009.0142
10. Güera D, Delp EJ (2018) Deepfake video detection using recurrent neural networks. In: 2018 15Th IEEE international conference on advanced video and signal based surveillance (AVSS), pp 1–6, https://doi.org/10.1109/AVSS.2018.8639163
11. Hu H, Li Y, Zhu Z, Zhou G (2018) Cnnauth: Continuous authentication via two-stream convolutional neural networks. In: 2018 IEEE International conference on networking, architecture and storage (NAS), pp 1–9
12. Khodabakhsh A, Ramachandra R, Raja K, Wasnik P, Busch C (2018) Fake face detection methods: Can they be generalized? In: 2018 International conference of the biometrics special interest group (BIOSIG), pp 1–6. https://doi.org/10.23919/BIOSIG.2018.8553251
13. Kim H, Garrido P, Tewari A, Xu W, Thies J, Niessner M, Pérez P, Richardt C, Zollhöfer M, Theobalt C (2018) Deep video portraits. ACM Trans Graph 37(4):163:1–163:14. https://doi.org/10.1145/3197517.3201283
14. Li Y, Chang M, 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. https://doi.org/10.1109/WIFS.2018.8630787
15. Li Y, Hu H, Zhu Z, Zhou G (2020) Scanet: Sensor-based continuous authentication with two-stream convolutional neural networks. ACM Trans Sen Netw 16(3):29:1–29:27. https://doi.org/10.1145/3397179
16. Li Y, Lyu S (2019) Exposing deepfake videos by detecting face warping artifacts. In: IEEE Conference on computer vision and pattern recognition workshops (CVPRW)
17. Li Y, Sun P, Qi H, Lyu S (2020) Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics. In: IEEE Conference on computer vision and patten recognition (CVPR). seattle, WA, United States
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. https://doi.org/10.1109/WACVW.2019.00020
19. Mo H, Chen B, Luo W (2018) Fake faces identification via convolutional neural network. In: Proceedings of the 6th ACM workshop on information hiding and multimedia security, IH&MMSec '18, pp 43–47, Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3206004.3206009
20. Nguyen HH, Yamagishi J, Echizen I (2019) Capsule-forensics: Using capsule networks to detect forged images and videos. In: ICASSP 2019 - 2019 IEEE International conference on acoustics, speech and signal processing (ICASSP), pp 2307–2311
21. 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
22. Rössler 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)
23. Schlegl T, Seeböck P, Waldstein SM, Schmidt-Erfurth U, Langs G (2017) Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In: IPMI
24. Suwajanakorn S, Seitz SM, Kemelmacher-Shlizerman I (2017) Synthesizing obama: Learning lip sync from audio. ACM Trans Graph 36(4):95:1–95:13. https://doi.org/10.1145/3072959.3073640
25. Thies J, Zollhofer M, Stamminger M, Theobalt C, NieBner M (2016) Face2face: Real-time face capture and reenactment of rgb videos, 2387–2395. https://doi.org/10.1109/CVPR.2016.262
26. Viola P, Jones M (2001) Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001, vol. 1, pp. I–I. https://doi.org/10.1109/CVPR.2001.990517
27. Wu J, Zhang C, Xue T, Freeman WT, Tenenbaum JB (2016) Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In: Proceedings of the 30th International conference on neural information processing systems, NIPS'16, pp 82–90, Curran Associates Inc., Red Hook, NY, USA
28. Yang LC, Chou SY, Yang Y (2017) Midinet: a convolutional generative adversarial network for symbolic-domain music generation. In: ISMIR
29. 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, https://doi.org/10.1109/ICASSP.2019.8683164

30. Yoo D, Kim N, Park S, Paek AS, Kweon IS (2016) Pixel-level domain transfer. In: ECCV
31. Zhang X, Karaman S, Chang S (2019) Detecting and simulating artifacts in gan fake images. In: 2019 IEEE International workshop on information forensics and security (WIFS), pp 1–6
32. 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 (CVPRW), pp 1831–1839. https://doi.org/10.1109/CVPRW.2017.229

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.