


REVIEW

A Survey on Deepfake Video Detection

Peipeng Yu¹  | Zhihua Xia^{2,3} | Jianwei Fei¹ | Yujiang Lu¹

¹Engineering Research Centre of Digital Forensics, Ministry of Education, School of Computer and Software, Jiangsu Engineering Centre of Network Monitoring, Jiangsu Collaborative Innovation Centre on Atmospheric Environment and Equipment Technology, Nanjing University of Information Science & Technology, Jiangsu Province, Nanjing, China

²College of Cyber Security, Jinan University, Guangzhou, China

³Engineering Research Center of Digital Forensics, Nanjing University of Information Science & Technology, Nanjing, China

Correspondence

Zhihua Xia, College of Cyber Security, Jinan University, Guangzhou, Guangdong Province, China
Province, 510632, China.

Email: xia_zhihua@163.com

Funding information

Collaborative Innovation Centre of Atmospheric Environment and Equipment Technology (CICAET) fund, China; Priority Academic Programme Development of Jiangsu Higher Education Institutions; '333' project of Jiangsu Province; Qinglan Project of Jiangsu Province; National Natural Science Foundation of China, Grant/Award Numbers: 61702276, 61772283, U1936118, 61601236, 61602253, 61672294, U1836208; National Key R&D Programme of China, Grant/Award Number: 2018YFB1003205; BK21+ programme from the Ministry of Education of Korea; Jiangsu Basic Research Programs-Natural Science Foundation, Grant/Award Number: BK20181407; Six peak talent project of Jiangsu Province, Grant/Award Number: R2016L13

Abstract

Recently, deepfake videos, generated by deep learning algorithms, have attracted widespread attention. Deepfake technology can be used to perform face manipulation with high realism. So far, there have been a large amount of deepfake videos circulating on the Internet, most of which target at celebrities or politicians. These videos are often used to damage the reputation of celebrities and guide public opinion, greatly threatening social stability. Although the deepfake algorithm itself has no attributes of good or evil, this technology has been widely used for negative purposes. To prevent it from threatening human society, a series of research have been launched, including developing detection methods and building large-scale benchmarks. This review aims to demonstrate the current research status of deepfake video detection, especially, generation process, several detection methods and existing benchmarks. It has been revealed that current detection methods are still insufficient to be applied in real scenes, and further research should pay more attention to the generalization and robustness.

1 | INTRODUCTION

The problem of face-manipulated videos has received widespread attention in the past two years, especially after the advent of deepfake technology that manipulates images and videos with deep learning tools. Deepfake algorithm can replace faces in the target video with faces in the source video using autoencoders or generative adversarial networks. With this technology, face-manipulated videos are exceedingly simple to generate on condition that one can access large amounts of data.

While deepfake technology could be used for positive purposes, such as film-making and virtual reality, it is still heavily applied for malicious uses [1–4]. As shown in Figure 1, a huge number of fake videos have been distributed on the Internet, most of which target at politicians and celebrities. The first deepfake content was a celebrity pornographic video

created by a Reddit user named *deepfakes* in 2017, which means that it is unavoidable for deepfake technology to be used for malicious uses since created. Soon after, FakeApp, FaceSwap and other deepfake-based applications appeared continuously. In June 2019, there is even a smart undressing app named Deepnude, resulting in a huge panic around the world. Except for damaging personal privacy, videos generated by these apps are increasingly applied to interfere in political campaigns and public opinion. The detection of deepfake content has become one of the hot issues for individuals, businesses and governments around the world.

With the increasing interest in deepfake technology, more and more related researches have been underway. The past two years have witnessed significant progress in developing new detection methods. To begin with, the number of video datasets built for deepfake detection tasks is growing. From small

This is an open access article under the terms of the Creative Commons Attribution-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited and no modifications or adaptations are made.

© 2021 The Authors. *IET Biometrics* published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology.



FIGURE 1 Video frames generated by deepfake algorithms. The first line shows the original video frames and the second line shows the corresponding video frames generated by deepfake methods

datasets (such as DeepFake-TIMIT [5] and UADFV [6]) in an early stage, to large-scale datasets (such as FaceForensic++ [7], Celeb-DF [8], DFDC [9] and DeeperForensic [10]), the number of datasets that can be used for training has increased. Furthermore, several research institutions are becoming aware of the dangers of deepfake videos and trying to promote related research. Recently, Amazon, Facebook and Microsoft have joined forces to host the deepfake detection challenge (DFDC) to build innovative technologies beneficial to detect deepfake videos. Also, SenseTime holds the deeperforensics challenge 2020 to solicit new ideas to advance the state-of-the-art in real-world face forgery detection. Owing to these reasons, several effective detection approaches have been proposed, which demonstrated excellent performance in forgery detection tasks.

Though several advances have been achieved, many critical issues for existing deepfake detection methods still need to be solved. With the continuous evolution of deepfake methods, generated videos become more and more realistic. In this case, traditional methods are probably not suitable for detecting manipulated videos generated by new deepfake algorithms [11]. It is significant to analyse and forecast the advanced development of deepfake-related research and improve corresponding detection approaches. In this review, we will focus on the existing detection scheme designed for deepfake videos, attempting to promote the development of deepfake video detection.

This article is organized as follows. In Section 2, we first introduce deepfake video generation algorithms proposed in recent years. Then, different types of detection approaches are described in Section 3. A list of datasets used in recent study is presented in Section 4. After that, a discussion of the state of deepfake video detection and its perspectives is carried out in Section 5. Finally, we conclude in Section 6.

2 | DEEPPFAKE VIDEO GENERATION

Since the first release of deepfake videos, new manipulation algorithms are proposed soon, most of which are based on generative networks. During these methods, deepfake algorithms can be used to create fake content to infringe on personal privacy, showing huge destructive effect on society. This

section will review the development of deepfake algorithms and then describe two types of deepfake algorithms.

2.1 | Development of deepfake technologies

Facemanipulation is not a new technology that appeared recently. The earliest attempt with facemanipulation in the literature can be found in the iconic 1865 portrait of the US President Abraham Lincoln. With the development of computer graphics technology, facemanipulation in digital images has become easily achievable [12–14]. Recent progress in the field of deep learning has fundamentally advanced the development of facemanipulation technology. According to different goals of face manipulation algorithms, existing deepfake algorithms could be divided into two categories: face swapping and face reenactment.

2.1.1 | Face swapping

Face swapping videos, swapping person identities in two videos, have attracted people's attention in recent years. Related researches have been established since 2017. In the study of Korshunova et al. [15], convolutional neural networks (CNNs) were trained to capture the appearance of target identity from an unstructured photo collection, which enables generating high-quality face-swapping images. However, time continuity is not considered, thus this approach cannot be applied for high-quality video generation. In the same year, Olszewski et al. [16] proposed a novel approach to generate videos with a single RGB image and a source video sequence. A deep generative network was used to infer perframe texture deformations of the target identity using source textures and the single target texture. Based on this method, the newly rendered face could be composited onto the source video, replacing the original face using the schema of [17]. In December 2017, the first face-swapping video generated by deepfake approach was posted by a Reddit user, bringing marvellous shock to the world. It is generally acknowledged that the inspiration of deepfake algorithms comes from [15], where CNNs were used to generate face-swapping images. After that, a wave of creating face-swapping videos was set off over the world, regardless of positive or negative purposes. Faceswap-GAN, an improved version of the original deepfake algorithms, was proposed in [18]. To generate more realistic faces, adversarial loss and perceptual loss were added to improve the performance of the autoencoder implemented by VGGFace [19]. Similarly, DeepFaceLab [20], an open-source deepfake generation framework, was designed for providing an imperative and easy-to-use pipeline for people without professional knowledge. In recent works, FaceShifter was proposed for occlusion aware face swapping with high fidelity [21]. Unlike previous face-swapping studies only using limited information from target images to synthesise faces, FaceShifter generates high-fidelity swapped faces by performing comprehensive integration of face attributes. Specifically, AEI-Net and HEAR-Net were leveraged to integrate face information and recover an

anomaly region, respectively. Experiments show its superior performance compared with existing face-swapping algorithms. Videos generated by recent deepfake approaches have been extremely realistic, hardly distinguished by human eyes.

2.1.2 | Face reenactment

Different from face-swapping technologies, face reenactment algorithms attempt to control people's expressions in videos, which means that attackers can generate videos manipulating someone to do something that does not exist. The first face reenactment algorithm could date back to 2006. Vlasic et al. [22] proposed to perform facial reenactment based on the face template, which was modified under different expression parameters. Most of the subsequent work is based on such schemes, where a parametric model is leveraged to adjust facial images. These methods could generate face images with high realism, but the obtained results often lack temporal coherence. In recent years, research on face reenactment has been further developed as computing ability increased. To perform monocular facial reenactment in real-time, Thies et al. [23] proposed Face2Face. In this study, a new global nonrigid model-based bundling approach was applied to reconstruct the facial features of target and source actors. At the same time, a subspace deformation transfer technique is designed to perform expression transfer between source and target actors. In addition to these contributions, this study also proposed a novel method in the synthesis of mouth regions, where the best matching image is retrieved from the target sequence. Compared to previous studies, Face2Face has already achieved quite remarkable performance. However, it cannot guarantee consistent head movements as only the migration of expressions is taken into account. Also, the synthesis of the mouth region is not satisfying, with coarse details of the mouth that are easily noticed by human eyes. With the development of deep learning techniques, these issues are gradually being noticed and addressed. It can be noticed that face videos synthesised by previous face reenactment algorithms have defects that they are inconsistent with voices. The work of Suwajanakorn et al. [24] supplemented this defect to a certain extent. They aimed to learn a sequence mapping from audio to video to manipulate actors to speak the same sentences as voice content. Features were extracted from voice sequence as the input of recurrent neural network (RNN), which outputs a sparse mouth shape corresponding to each frame of video output. The textures of mouth are further synthesized and merged into the original video. A better improvement is achieved by Fried et al. [25]. They performed talking-head video editing and changing speech words by using designed neural face rendering method. To perform face reenactment with better performance, Kim et al. [26] proposed a new method for photorealistic reanimation of portrait videos. The proposed generative neural network with a novel space-time architecture is used to transform coarse face model renderings into full photorealistic portrait video output. The major contribution of this study is designing a new spatiotemporal encoding as conditional input for video synthesis, resulting in synthesised

videos with a high degree of spatiotemporal continuity. Compared to Face2Face, this work can not only migrate facial expressions, but also head pose, gaze direction and blinking movements, compensating for the inaccurate head pose in the Face2Face algorithm. Except for this study, Thies et al. [27] has also made further optimisations to address problems existing in Face2Face. Neuraltexture, incorporating Face2Face and neural networks for texture extraction based on Face2Face, compensating for Face2Face's blurred texture in the mouth region.

2.2 | General process of deepfake video generation

In this part, we will briefly describe the generation process of two types of deepfake videos.

2.2.1 | Face swapping

To generate a face-swapping video, all frames of the target video have to be processed using generative method. Figure 2 shows the general generation process of face-swapping videos. Obviously, the deepfake algorithm, which implements faceswapping while preserves the source expressions, is the core part of video generation. The deepfake algorithms used in faceswapping are mostly developed based on autoencoder, which is widely used for data reconstruction tasks. Autoencoder is composed of two components: an encoder and a decoder. Latent features are first extracted from the image by the encoder, and then inputted to the decoder to reconstruct the original image. In the deepfake algorithm, two autoencoders are trained to swap faces between source video frames and target video frames. As shown in Figure 3, during the training process, two encoders with the same weights are trained to extract common features in source and target faces. Then, features extracted are inputted to two decoders to reconstruct faces, respectively. It is worth noting that decoder A is only trained with faces of A while decoder B is only trained with faces of B. When the training process is complete, a latent face generated from face A will be passed to the decoder B. Decoder B would try to reconstruct face B from feature relative to face A. If the autoencoder is trained well, the latent space will represent facial expressions. In other words, the face generated by decoder B will have the same expression as face A.

2.2.2 | Face reenactment

The face reenactment task aims to perform the migration of facial expressions. In order to better demonstrate this kind of scheme, we directly use the scheme in [26] as an example to introduce. Figure 4 shows the general process of performing face reenactment. First, the low-dimensional parameter representation of the source and target videos is obtained using a monocular face reconstruction method. Furthermore, head pose and expression could be transferred to the parameter

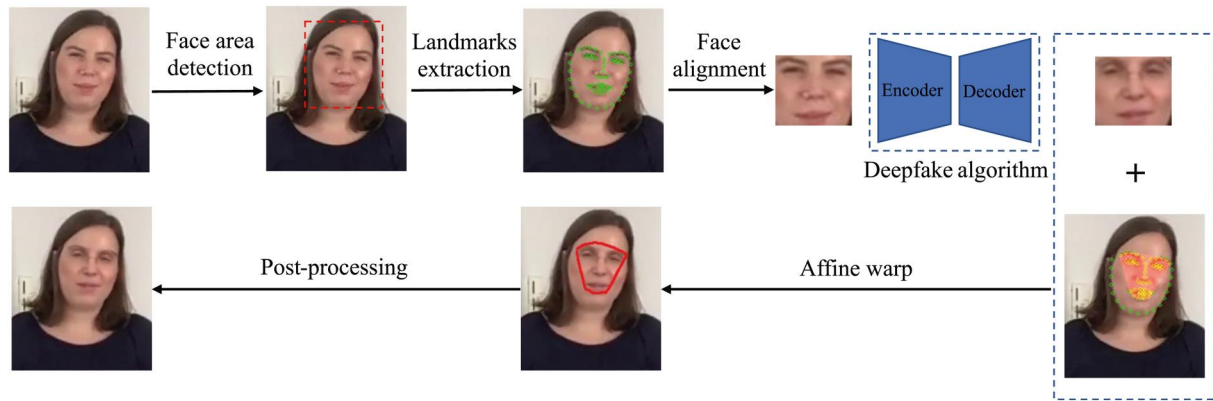


FIGURE 2 The generating process of face-swapping video frames. The face area is first detected in each video frame. Then, facial landmarks are extracted to perform face alignment. After that, the deepfake algorithm (autoencoder or GAN) is applied to generate a synthetic face by inputting the face-aligned image. To reduce artefacts caused by blending, the landmarks of the left and right eyebrows and the bottom mouth are used to generate a specific mask. In this way, after blending the synthetic face to the original image, only the content inside the mask is retained. Finally, to further make the generated image realistic, a postprocessing operation is supplemented to process the generated image. Specifically, Gaussian blur is applied to the boundary of the mask while the colour correction algorithm is applied to ensure the consistency of the synthetic face and background image

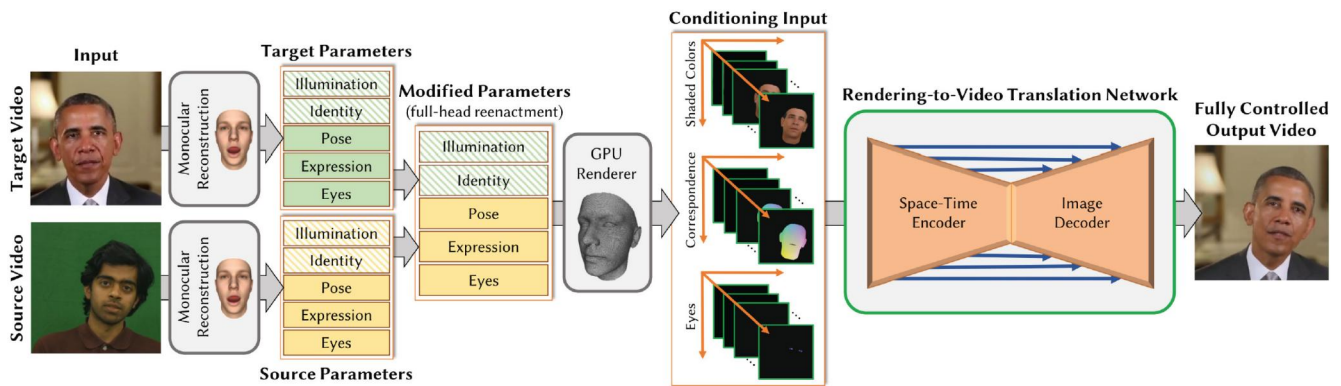


FIGURE 3 Generating process of face reenactment videos [26]. First, monocular face reconstruction is performed on the source face and the target face to obtain their respective face parameters. After that, parameters are modified by preserving parameters of illumination and identify while changing parameters of pose, expression and eye gaze. Synthetic images are then generated using modified parameters. Finally, rendering-to-video translation network is applied to generate face reenactment videos

space. To perform face reenactment, scene illumination and identity parameters are preserved while head pose, expression and eye gaze parameters are changed. After that, synthetic images of the target actor are regenerated based on the modified parameters. These images are then served as the conditional input of our new renderingvideo conversion network, which is then trained to convert the synthesized input into a realistic output. To obtain a complete video with better time consistency, the conditioning space-time volumes are fed into the network in a sliding window fashion. In this way, face reenactment video can be obtained.

3 | DEEFAKE VIDEO DETECTION

Deepfake videos are increasingly harmful to personal privacy and social security. Various methods have been proposed to detect manipulated videos. Early attempts mainly focused on inconsistent features caused by the face synthesis process while

current detection methods mostly target at fundamental features. As shown in Table 1, these methods fall into five categories based on the features they use. To begin with, detection based on general neural networks is commonly used in literature, where deepfake detection task is considered as regular classification tasks. Temporal consistency features are also exploited to detect discontinuities between adjacent frames of fake video. To find distinguishable features, visual artefacts generated in blending process are exploited in detection tasks. Recently proposed approaches focus on more fundamental features, where camera fingerprint and biological signal-based schemes show great potential in detection tasks. In the following sections, we will review detection methods mentioned above.

3.1 | General-network-based methods

Recent advances in image classification have been applied to improve the detection of deepfake videos. In this method, face

images extracted from the detected video are used to train the detection network. Then, the trained network is applied to make predictions for all frames of this video. The predictions are finally calculated by averaging or voting strategy. Consequently, the detection accuracy is highly dependent on the neural networks, without the need to exploit specific distinguishable features. In this section, we divide existing network-based methods into two types: transfer learning-based methods and detection approaches based on specially designed networks.

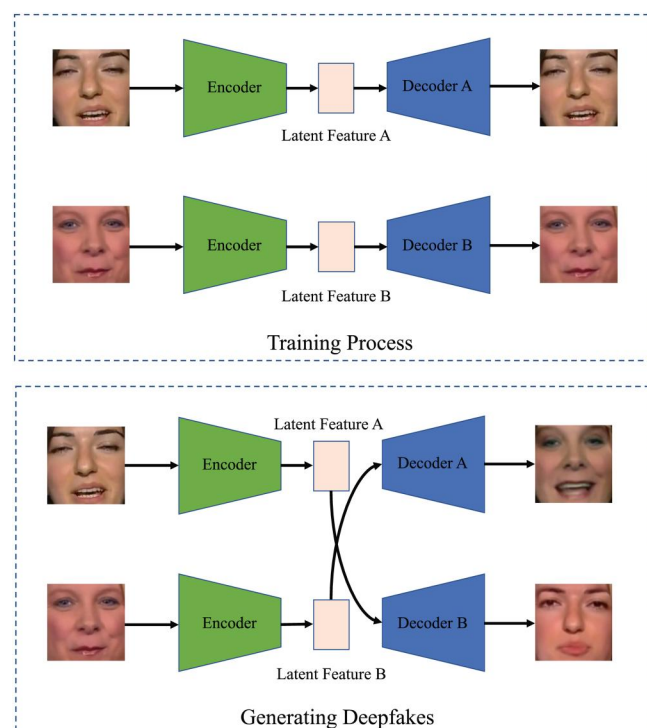


FIGURE 4 Autoencoders used for faceswapping. Figure top shows the training process of the two autoencoders. When generating deepfake faces, decoders are swapped as the figure below shows

TABLE 1 Classification for existing detection methods

Methods	Description
Generalnetwork-based methods	In this method, detection is regarded as a frame-level classification task which is finished by CNNs
Temporalconsistency-based methods	Deepfake videos are found to exist inconsistencies between adjacent frames due to the defects of the forgery algorithm. Thus RNN is applied to detect such inconsistencies
Visualartefacts-based methods	The blending operation in generation process would cause intrinsic image discrepancies in the blending boundaries. CNN-based methods are used to identify these artefacts
Camerafingerprints-based methods	Due to specific generation process, devices leave different traces in the captured images. At the same time, faces and background images are acknowledged to come from different devices. Thus, detection task can be completed by using these traces
Biologicalsignals-based methods	GAN is hard to understand hidden biological signals of faces, making it difficult to synthesize human faces with reasonable behaviour. Based on this observation, biological signals are extracted to detect deepfake videos

3.1.1 | Transfer learning

Network-based detection methods should be the earliest method introduced for detection tasks. Shortly after the appearance of the first deepfake video, some early detection algorithms were proposed, mainly based on existing networks that performed well in image classification tasks. Transfer learning strategy could be easily found in the early studies. Combining steganalysis features and deep learning features, Zhou et al. [28] put forward a two-stream network for face tampering detection. Likewise, in [7], Rossler et al. evaluated XceptionNet [29] on the FaceForensic++ dataset, outperforming all other networks in detecting fakes. During DFDC, similar detection methods were used. In [30], two existing models were tested to provide a performance baseline: A small DNN (composed of six convolutional layers and a fully connected layer) and an existing XceptionNet. Early results showed that the best method (XceptionNet) provides 93.0% precision. Bonettini et al. [31] studied the ensemble of different trained CNN models, showing that the ensemble of CNNs can achieve promising results in deepfake detections. However, such network-based algorithms are prone to overfitting [32], so researchers attempted to exploit intrinsic differences between real and fake videos through preprocessing. Some preprocessing methods, such as optical flow calculation [33], had been proved to be useful to exploit interframe dissimilarities in network-based methods.

3.1.2 | Specially designed networks

With the advent of large-scale datasets and the development of detection algorithms, more attention is attracted to improve the generalization of detection algorithms. Nguyen et al. [34] introduced a capsule network to improve the performance of detection networks. As illustrated in Figure 5, face images are first fed into the pretrained VGG-19 network [35]. Extracted features are then inputted into the proposed capsule network, which includes several primary capsules and two output capsules. Agreement between the features

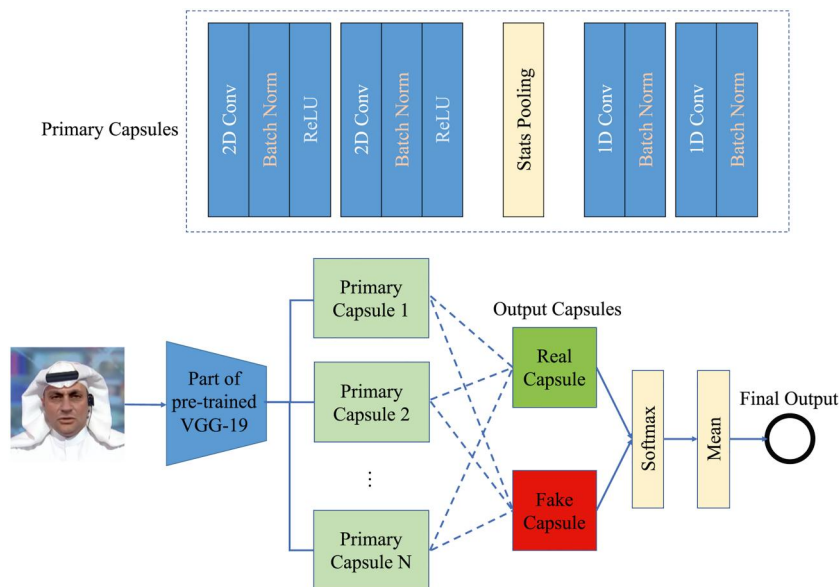


FIGURE 5 Capsule-forensics architecture. Pretrained VGG-19 is first used to extract features from face images. Features are further input into proposed capsules, which include several primary capsules and two output capsules. Agreement between primary capsules and output capsules is calculated by a dynamic routing algorithm. Finally, output of capsules is mapped to probabilistic values

extracted by the primary capsules is dynamically calculated by a dynamic routing algorithm and the results are finally routed to the appropriate output capsule. Visualization of latent features extracted indicated that the combination of capsule networks and dynamic routing algorithm is effective for detecting manipulations. However, the capsule network performed terribly when encountering unknown deepfake videos [8], proving that capsule networks still need further improvement to detect high-fidelity videos. To explore the mesoscopic properties of images, Afchar et al. [36] also proposed a CNN, namely MesoInception-4, consisted of a variant of inception modules introduced in [37]. Their proposed approach achieved 98.4% accuracy using a private database. Moreover, this approach is also tested using unseen datasets in recent study [7, 8, 30, 38], proving to be a robust approach in deepfake detection tasks. Although these methods achieved excellent results on various datasets, the reasons behind good performance are still unknown. In fact, deeper networks tend to achieve better results than shallower network in various areas. The reason for the good performance may simply be that designed networks are deep enough. Compared with traditional learning-based methods, Wang et al. [39] pay more attention to neuron coverage and interactions rather than the design of specific network structures. The *FakeSpotter* they proposed uses hierarchical neuron behaviour as a feature, showing high robustness against four common perturbation attacks. This research provided a new insight for detecting fakes.

3.1.3 | Summary

The disadvantage of network-based methods is that such methods tend to overfit on specific datasets. In this type of method, although adjustment and optimization of model

structure often affect the abstraction degree of features, it still lacks sufficient relevance for the task of deepfake detection. Therefore, the current direction of such work is gradually changing. On the one hand, by adding additional components to the model, the model can be constrained to learn heuristic features [40]. In this case, importance of model architecture is greatly reduced while additional components play a greater role. This is exactly the difference between deepfake detection tasks and general computer vision tasks. On the other hand, more and more network-based methods have begun to introduce multitask learning, that is, not only to classify real and fake faces, but also to generate pixel-level tampering masks. In [41], using a semi-supervised learning strategy, Nguyen et al. designed a multitask learning framework to simultaneously detect manipulated content and locate the manipulated regions. In such schemes, however, supervised multitask learning is only a complementary implementation, which does not necessarily improve the final detection performance. Further improvement was achieved by using attention mechanisms. Dang et al. [42] utilized an attention mechanism to process feature maps for the classification. The proposed approach showed excellent performance both in deepfake detection and forgery location, achieving state-of-the-art performance compared to previous solutions. Their approach demonstrates the importance of attention mechanisms. Likewise, in [43], Tarasiou et al. designed a lightweight architecture for extracting local image features and a multitask training scheme for forgery localization. In this way, the forgery location process provides evidence for judgement while ensuring detection accuracy, promoting the practical use of detection algorithm. It is worth mentioning that some basic directions in computer vision, such as anomaly detection, semantic segmentation and metric learning, are making more and more important contributions in this field.

3.2 | Temporal-consistency-based methods

Time continuity is a unique feature for videos. Unlike images, video is a sequence composed of multiple frames, where adjacent frames have a strong correlation and continuity. When video frames are manipulated, the correlation between adjacent frames will be destroyed due to defects of deepfake algorithms, specifically expressed in the shift of face position and video flickering. According to this phenomenon, researchers have proposed several detection approaches. We will first introduce the original CNN-RNN architecture and then demonstrate its improvement in these years.

3.2.1 | CNN-RNN

Considering the time continuity in videos, Guera et al. [44] first proposed to use RNN to detect deepfake videos. In their work, autoencoder was found to be completely unaware of previously generated faces because faces were generated frame-by-frame. This lack of temporal awareness results in multiple anomalies, which are crucial evidence for deepfake detection. To check the continuity between adjacent frames, an end-to-end trainable recurrent deepfake video detection system was proposed. As Figure 6 shows, the proposed system is mainly composed of a convolutional long short-term memory (LSTM) structure for processing frame sequences. Two essential components are used in a convolutional LSTM structure, where CNN is used for frame feature extraction and LSTM is used for temporal sequence analysis. Specifically, a pretrained inceptionV3 [45] is adapted to output a deep representation for each frame. The 2048-dimensional feature vectors extracted by the last pooling layers are applied as the sequential LSTM input, characterizing the continuity between image sequences. Finally, a fully connected layer and a softmax layer are added to compute forgery probabilities of the frame sequence tested. The experiments on a self-made dataset showed that the algorithm can accurately make predictions even when the length of a video is less than 2 s. Although this research did not show its superiority since there were no large-scale datasets at the time, several articles after were inspired by

this article, which promoted the development of detection methods based on temporal consistency.

3.2.2 | Improvement

After the time-based detection method showed its effectiveness, many related studies were proposed. In [46], Sabir et al. utilized the temporal information present in the video stream to detect deepfake videos. Similar to [44], an end-to-end model is built, where CNN is also involved in the follow-up training. Meanwhile, face alignment based on facial landmarks and spatial transformer network is applied to further improve the performance of the algorithm. Even though such solutions guarantee high accuracy in videos with high quality, they do not perform well on low-quality video when the continuity between adjacent frames is disrupted by video compression operations. To solve this problem, a CNN-RNN framework based on automatic weighting mechanisms was proposed by Montserrat et al. [47]. Considering that the face qualities of some frames are not high, an automatic weighting mechanism was proposed to emphasize the most reliable regions when making a video-level prediction. Experiments showed that combining CNN and RNN achieves high detection accuracies on the DFDC dataset. Except for the robustness of algorithms, generalization ability is also essential for forgery detection tasks. Zhao et al. [48] used optical flow to capture the obvious differences of facial expressions between adjacent frames. However, these studies did not show strong generalization or robustness. To solve this problem, Wu et al. [49] proposed a novel manipulation detection framework, named SSTNet, exploiting both low-level artefacts and temporal discrepancies. Another study proposed by Masi et al. [50] obtained good generalization on multiple datasets. In their research, a two-branch recurrent network is applied to propagate the original information while suppresses the face content. Multiband frequencies are amplified using a Laplacian of Gaussian as a bottleneck layer. Inspired by [51], a new loss function is designed for better isolating manipulated face. The experimental results on several datasets show the excellent generalization performance of the detection algorithm. Nevertheless, time-based detection schemes still have much

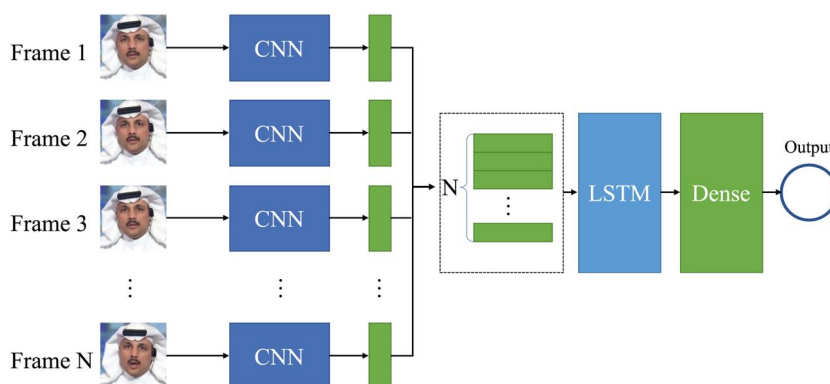


FIGURE 6 Overview of detection method based on CNN-LSTM. The backbone CNN model is first used to extract features of each face image in successive frames. Output features are then merged and used as input to the LSTM network, which processes the time series features to obtain the probability value of whether the video clip is true or false

room for improvement in generalization performance [47]. Screen switching and unknown data are still problems that need to be solved for time-based detection approaches.

3.2.3 | Summary

Compared with general-network-based approaches, temporal-consistency-based detection methods consider the continuity between adjacent frames, thereby improving the detection performance. However, many models tend to destroy the spatial structure of original frames when extracting temporal features while the motivation for designing such methods is precisely to extract the inconsistency of spatial features in the temporal domain. CNN-RNN architectures pool the intra-frame features into vectors [44, 46], thus cannot capture spatial features while detecting temporal consistency. Although structures such as 3DCNN can avoid destroying spatial features, the excessive parameters make it easier to overfit on a specific dataset.

3.3 | Visualartefacts-based methods

In most existing deepfake methods, the generated face has to be blended into an existing background image, causing exist intrinsic image discrepancies on the blending boundaries. As shown in Figure 7, faces and background images come from different source images, giving rise to the abnormal behaviour of the synthetic image, such as boundary anomaly and inconsistent brightness. These visual artefacts make deepfake videos fundamentally detectable. In this section, three main visual artefacts would be introduced.

3.3.1 | Face warping artefacts

Based on the observations with inconsistency between faces and background, a new deep learning-based method was proposed by Li and Lyu [38]. Face warping artefacts generated by blending process were used to detect fake videos. As shown in Figure 2, synthetic faces have undergone an affine transform to match the poses of the target faces. In this case, there would be an obvious colour difference and resolution inconsistency between the internal face and background areas. Since the purpose here is to detect inconsistency between the face region



FIGURE 7 Video frames with visual artefacts. Deepfake generated image shows colour difference and resolution inconsistency because of the lack of postprocess

and background area, the negative samples are generated by a simplified process, where the face undergoes an affine warp back to the source image directly after smoothed. To generate more realistic negative examples, a convex polygon shape is used based on the face landmarks of eye browns and the bottom of the mouth. Also, colour information is also randomly changed to enlarge the training diversity. After that, four CNN models—VGG16, ResNet50, ResNet101 and ResNet152 were trained in this study. Evaluated on several datasets of available deepfake videos, this method demonstrated effectiveness in practice. Compared with previous methods, this study focuses on the visual artefacts caused by affine transformation. At the same time, due to no additional negative samples to participate, this algorithm does not need to fit the sample distribution of deepfake videos, greatly increasing the generalization of the algorithm [8].

3.3.2 | Blending boundary

Further improvements were achieved in [32]. Li et al. proposed a novel image representation, namely face X-ray, which was exploited to observe whether the input image can be decomposed into the foreground face and the background. Specifically, the blending boundary between the foreground manipulated face and the background was defined as face X-ray. Compared with Li and Lyu [38], this study targeted at the blending boundary that is universally introduced in image blending, thus showing great performance when tested in various datasets. Except for proposing face X-ray, this research designs the generation process of negative samples by using positive samples particularly. Thus, the algorithm does not need to consider face manipulation in the deepfake video, but only focuses on the difference between background and foreground faces, thereby enhancing the generalization of the proposed algorithm. However, due to excessive focus on the blending boundary, this scheme is not resistant to fully synthesized images.

3.3.3 | Head pose inconsistency

Another interesting study comes from [52]. Observing that deepfake videos were created by splicing a synthesised face into the original image, Yang et al. proposed a new detection method based on 3D head poses. They argued that current generative neural networks could not guarantee landmark matching, causing that estimated 3D landmarks on the face-manipulated area were different from 3D landmarks estimated from the whole face area. In this method, the rotation matrix estimated using facial landmarks from the whole face and the one estimated using only landmarks in the central region are calculated to analyse the similarity between two pose vectors. Although the experiment confirmed the difference between real and fake pose vectors, this study was built based on specific features existing in a self-made dataset which was generated by relatively basic version of the deepfake algorithm.

Thus, this method is not effective for detecting a new version of deepfake videos as deepfake algorithms evolve [8].

3.3.4 | Summary

Visual-artefacts-based methods often obtain better generalization performance because they target more general artefacts existing in most deepfake contents. However, these algorithms can only detect specific forgery traces due to paying more attention to specific artefacts. With the progress of deepfake algorithms, these artefacts are gradually disappearing. Nevertheless, visual artefacts-based approaches obtain better performance in the latest version of deepfake video datasets. Such schemes still have high potential in deepfake detection tasks. Researches should be established to exploit more intrinsic features.

3.4 | Camera-fingerprints-based methods

Camera fingerprints are a kind of noise with very weak energy, which plays an important role in forensic fields, especially source identification tasks. In general, camera-fingerprints-based approaches have gone through three processes: the photo response nonuniformity (PRNU) patterns, noiseprint and recent video noise pattern. We will introduce its development in the following content.

3.4.1 | PRNU patterns

The detection based on camera fingerprints originated from image forensics. Observing that devices will leave different traces in the captured images, Lukas et al. [53] proposed PRNU noise, which can be used in camera identification tasks. PRNU arises due to the different sensitivities of the pixels to light caused by the inhomogeneity of the silicon wafer and imperfections in the sensor manufacturing process. Because of uniqueness and stability, the PRNU pattern is regarded as a device fingerprint, which can be used to carry out many forensic tasks [54–56]. Based on these findings, Koopman et al. [57] first proposed to use PRNU to detect deepfake videos. PRNU patterns are verified to be effective on a small dataset. However, in [58], the PRNU-based classifier achieves much lower accuracy when tested in GAN-generated datasets. More researches should be performed to verify the effectiveness of the PRNU pattern in deepfake detection tasks.

3.4.2 | Noiseprint

In fact, PRNU-based methods can only extract device-related features while suppressing other camera artefacts existing in the image generated process. Traces generated during the digital image acquisition process are composed of several noises. Inside the camera, the image undergoes operations such

as interpolation and gamma correction. Outside the camera, the image could also be compressed or enhanced, which will leave many traces in the final image. Thus, each image has its unique traces, namely noise residuals, which can be used to identify its source camera. Following this direction, Cozzolino et al. introduced a CNN-based camera fingerprint named noiseprint in [59]. To remove scene content and enhance camera model-related artefacts, a siamese network was trained using images coming from different camera models. In this siamese network, a fully convolutional network, proposed in [60], was first introduced to extract the noise pattern of images. Pairs of images from the same or different camera models were used to train the siamese network. At the end of the training process, CNN used in the siamese network could be used to extract the corresponding noiseprint from the input image, displaying enhanced camera model artefacts. This work provides new ideas for fingerprint noise extraction tasks, further promoting the development of image forensic area.

3.4.3 | Video noise pattern

After introducing the concept of noiseprint, Cozzolino extended these findings to the video forensic area [59]. Except for source identification, noiseprint was also adopted for forgery detection and localization. Considering that in a manipulated video, the manipulated region was generated differently from the background region and therefore carries different noises, they argued that forgery detection could be finished by using video noiseprint. As shown in Figure 8, noiseprints are extracted frame-by-frame, which are then averaged to indicate the noise contained in the video. Face and background regions are then split to calculate the similarity. Similar to [59], the spatial co-occurrences matrix of the extracted noiseprint is used to further calculate the Mahalanobis distance between the face region and the reference, which is then used as the manipulation score. The algorithm showed good detection performance on the FaceForensic++ dataset, even though the noise extraction network had not been trained on it. However, since the noiseprints extracted from frames are averaged to represent video noiseprint, the calculation of video noiseprint will be interfered if the video has a large motion. In this way, though noiseprint has shown its effectiveness in image manipulation, its using strategy in video forensic area still needs further improvement.

3.4.4 | Summary

Camera fingerprints have been proved to be effective in deepfake detection tasks. However, accurate estimation of camera fingerprints requires a large number of images captured by different types of cameras. Thus, there would be a decrease in accuracy when detecting images captured by unknown cameras. On the other hand, camera-fingerprint-based methods are not robust to simple image postprocessing such as compression, noise and blur. Since GAN images are generated

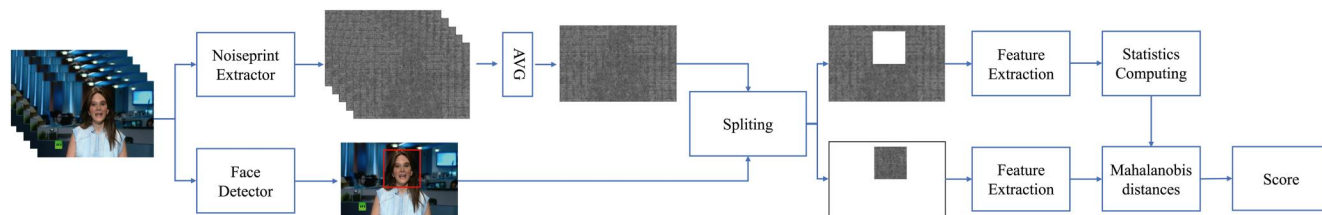


FIGURE 8 Scheme used for deepfake detection. Noiseprints are first extracted on a sufficient number of video frames. Then, extracted noiseprints are averaged to represent the video noiseprint. Divided by the face detector, the video noiseprint is then split into face region and background region. After that, the algorithm extract features of background region and calculate the statistical information. Finally, the Mahalanobis distance between features of face and background area is calculated to obtain the final heat map

without any image capture process, there is no camera fingerprint in the output image, so that the camerafingerprint-based methods are very suitable for detecting images generated by GANs. However, recent work shows that images can also be generated by simulating camera fingerprints [61], thus deceiving detection methods that rely on camera fingerprints. Recent research also proved that noise pattern could be erased by neural networks [62]. In this way, existing camerafingerprint-based methods should increase robustness to resist such attacks.

3.5 | Biological-signals-based methods

Detection based on biological signals is an interesting scheme that emerged in recent years. The core observation is that even though GAN is able to generate faces with high realism, the naturally hidden biological signals are still not easily replicate, making it difficult to synthesize human faces with reasonable behaviour [63]. Taking advantage of this abnormal behaviour, several studies have been proposed. In this section, we will introduce two approaches based on biological signals: blinking frequency-based and heart rate-based detection approaches.

3.5.1 | Eye blinking

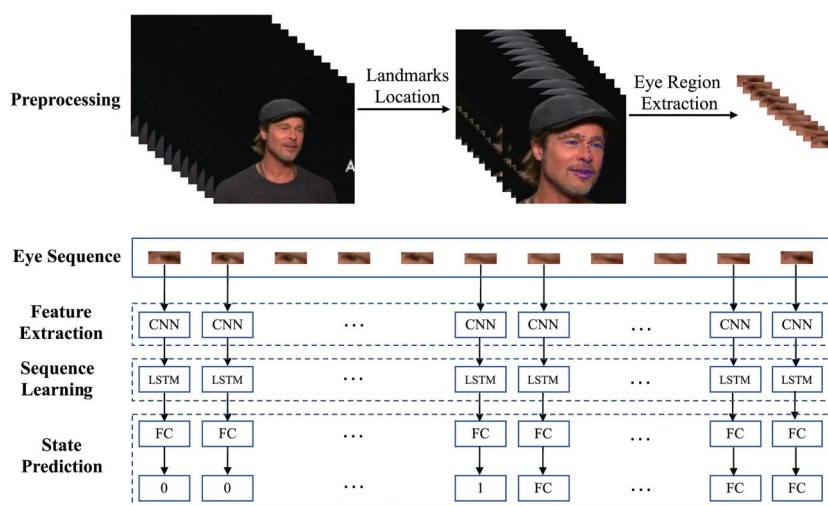
Abnormalities with blink frequency were earlier identified as discriminable features in deepfake detection tasks [6]. This could be attributed to the fact that deepfake algorithms train models using a large number of face images obtained online. Most of the images show people with their eyes open, causing that a closed-eye view is difficult to generate in a manipulated video. Based on this finding, a deep neural network model, known as long-term recurrent CNN (LRCN) [64], was introduced to distinguish open and close eye states. To calculate blink frequency, surrounding rectangular regions of eyes are cropped into a new sequence of input frames after face alignment. Then, the cropped sequences are passed into the LRCN model to capture temporal dependencies. As shown in Figure 9, the feature extraction module is first used to extract discriminative features from the input eye region by a CNN based on VGG16 framework. The output of feature extraction is then fed into sequence learning, implemented with an RNN

model. In the final state prediction stage, a fully connected layer is added to calculate the probability of eye open and closed states, which is then used to calculate blink frequency. This method is evaluated over self-made datasets, showing promising performance on detecting videos generated with deepfake methods. However, forgery algorithms can easily generate videos with reasonable blinking frequency as long as enough closed-eye images are added to the training set. Due to excessive attention to abnormal blinking frequency, this method is no longer applicable for current deepfake detection tasks after the problem of blink frequency is solved.

3.5.2 | Heart rate

Except for blink frequency, heart rate was also found the difference between real and manipulated videos. Previous literature had proved that colour changes of skin in the video could be applied to infer heart rate [65–67]. Based on these findings, a detector based on biological signals named FakeCatcher was designed to detect deepfake videos [63]. Specifically, remote photoplethysmography (rPPG or iPPG) was used to extract heart rate signals according to subtle changes of colour and motion in RGB videos [55, 68]. Experiments validated that spatial coherence and temporal consistency of such signals are not wellpreserved in deepfake videos. Following statistical analysis, a robust synthetic video classifier was developed based on physiological changes. Results verified that FakeCatcher has a high detection accuracy for deepfake videos, even for low-resolution or low-quality videos. Similarly, Fernandes et al. [69] proposed to use neural ordinary differential equations [70] to predict the heart rate of deepfake videos. A large difference was shown between original videos and deepfake videos when heart rate prediction was performed separately. However, this work only performed heart rate prediction of deepfake videos while lacked further experiments of deepfake detection. A large number of works have been carried out based on biological signals. A recently proposed approach, named DeepRhythm [71], utilized a dual-spatial-temporal attention mechanism to monitor the heartbeat rhythms, proving to generalize well over different datasets. Likewise, DeepFakesON-Phys [72] predicts the heart rate through changes in skin colour, thereby considering the detection of deepfake videos.

FIGURE 9 Overview of LRCN method. Eye sequences are first extracted by preprocessing module, then inputted to feature extraction module to generate feature sequences. Sequence learning module is then applied to analysis time-related sequences. Finally FC layer is added to make state prediction, calculating the blinking rate



3.5.3 | Summary

Although biologicalsignals-based detection approaches have shown good performance on various datasets, the natural flaw of this kind of method is that the detection process cannot be performed in an end-to-end way. Also, the information reflected by biological signal is seriously affected by video quality, so there are natural flaws and limited application range for biologicalsignals-based approaches.

4 | DATASETS AND PERFORMANCE EVALUATION

Since current deep learning-based methods are highly dependent on large-scale data, building high-quality data sets reflects importance. As deepfake algorithms evolve, new datasets should be built to develop advanced algorithms to counter new manipulation methods. In this section, we will describe the most commonly used datasets shown in Table 3, and briefly introduce their characteristics. Detection performance on these datasets will also be introduced.

Due to the lack of publicly available datasets, several self-made datasets were built to verify the effectiveness of proposed algorithms in the early literature. In [73], Deepfake-TIMIT, composed of 620 deepfake videos for 16 pairs of subjects, was proposed to evaluate several baseline face swap detection algorithms. Also, in [6], UADFV dataset was collected to detect eye blinking rate in the videos. The dataset consists of 49 original videos from YouTube and 49 deepfake videos generated by FakeApp, with a typical resolution of 294×500 pixels and an average time of 11.14 s. These self-made datasets greatly promoted the development of deepfake detection algorithms in the early stage. As shown in Table 2, detection algorithms perform well on these datasets. However, fake videos generated in these datasets are often targeted at specific detection algorithms and the quality of these videos is not enough for current detection tasks. The first large-scale dataset used for deepfake detection is

FaceForensic++, introduced by Rossler et al. [7]. The dataset contains 1000 original videos and 4000 manipulated videos generated by four different forgery methods. Specifically, these methods contain DeepFake, FaceSwap, Face2Face and NeuralTexture, where the first two are used to swap faces and the latter two are used for expression manipulation. Videos of three different compressed levels are provided to develop robust detection methods. Several studies have verified the effectiveness of proposed methods using this dataset. However, the forgery method used to generate negative samples in the dataset is relatively backward compared with current deepfake algorithms, causing a multitude of visual artefacts in generated forged videos. As shown in Table 4, the detection accuracy of various detection schemes has reached even more than 99% on the FaceForensic++ dataset. Although FaceForensic++ has made great contributions to the development of deepfake detection, it is inconsistent with the current development status of deepfake research, thus cannot be used to verify the performance of current detection algorithms.

To further simulate the realistic scene, datasets generated by novel deepfake algorithms are proposed. Google and Jigsaw proposed deepfake detection dataset [75], a large-scale dataset built for deepfake detection. In this dataset, 3000 deepfake videos are created by 28 actors in various scenes, which are more realistic than FaceForensic++. After commercial companies took part in deepfake detection research, DFDC was held to promote the development of deepfake detection. During the challenge, two datasets were introduced: DFDC-preview dataset [30] and DFDC dataset [9]. The DFDC-preview dataset is built by two different deepfake approaches, composed of 1131 original videos and 4113 corresponding deepfake videos. DFDC dataset is a much larger dataset used for competition in Kaggle, consisted of over 470 GB of videos (pristine and manipulated). It is worth noticing that in order to promote the practical application of the deepfake detection algorithm, DFDC is more random in the data collection, bringing more visual variability. Related research on the DFDC dataset and the top-3 detection

TABLE 2 Detection performance on self-made datasets

Study	Method	Dataset	Performance	FLOPs
Zhou et al. [28]	A two-stream network	SwapMe and FaceSwap dataset	0.927 (AUC)	>5.73
Guera et al. [44]	CNN + LSTM	Self-made dataset	97.1% (ACC)	>5.73
Yang et al. [52]	3D head poses	UADFV	0.974 (AUC)	-
Li et al. [6]	Eyeblink + LRCN	Self-made dataset	0.99 (AUC)	15.5
Ciftci et al. [63]	Biological signals	Self-made deep fakes dataset	91.07% (ACC)	-
Afchar et al. [36]	MesoInception-4	Meso-data(frame-level)	91.70% (ACC)	0.5
		Meso-data(video-level)	98.4% (ACC)	
Nguyen et al. [34]	A capsule network	Meso-data(frame-level)	95.93% (ACC)	>7.72
		Meso-data(video-level)	99.23% (ACC)	
Li and Lyu [38]	Face warping artefacts + CNN	UADFV	0.974 (AUC)	4.12
		Deepfake-TIMIT(LQ)	0.999 (AUC)	
		Deepfake-TIMIT(HQ)	0.932 (AUC)	
Li et al. [74]	Patch an pair CNN	Mesonet-data	0.979 (AUC)	1.82
		Deepfake-TIMIT	1.0 (AUC)	

TABLE 3 List of datasets including video manipulations

Dataset	Release date	Real/fake	Source
UADFV [6]	2018.11	49/49	YouTube
Deepfake-TIMIT [5]	2018.12	-/620	YouTube
FaceForensics++ [7]	2019.01	1000/4000	YouTube
Google DFD [75]	2019.09	363/3068	Actors
DFDC-preview [30]	2019.10	1131/4119	Actors
DFDC [9]	2019.10	23,654/104,500	Actors
Celeb-DF [8]	2019.11	890/5639	YouTube
DeeperForensics [10]	2020.1	10,000/50,000	Actors

scheme of DFDC are shown in Table 5. It is believed that DFDC dataset would bring more contributions to the development of deepfake detection tasks.

Although the scale of the current deepfake video dataset has been able to meet the needs of detection algorithm, videos in these datasets have obvious visual artefacts, which are not in line with the current status of existing deepfake approaches. To solve this problem, Li et al. [8] introduced Celeb-DF dataset, generated by an improved deepfake approach. Problems existing in the early version of deepfake videos, such as temporal flickering and low resolution of synthesized faces, are improved in this dataset. The dataset is comprised of 590 real videos and 5639 deepfake videos, satisfying the need for model training. Experimental results shown in literature (shown in Table 6) prove that Celeb-DF is currently the most challenging dataset, where the detection accuracy of various methods on Celeb-DF is lower than that of other datasets.

Another large-scale benchmark, composed of 50,000 original videos and 10,000 manipulated videos, has been built in [10]. DF-VAE, a new conditional autoencoder, is applied to generate deepfake faces with a higher realism rating. Studies using DeeperForensics demonstrates that the quality of the generated video is significantly better than that of the existing dataset.

5 | DISCUSSION

Deepfake videos appeared in people's attention in the past two years, posing a serious threat to social security. To this end, researchers have carried out a large number of research and achieved remarkable advances. Recent detection algorithms achieve almost 100% detection accuracy in the earlier deepfake dataset. However, the accuracy of existing detection algorithms is not ideal in recently built datasets. In the recent DFDC competition, the average accuracy of detection approaches proposed in the entire competition is only 65.18%, proving that current detection approaches are still far from meeting the needs of practical scenes. At the same time, current research tends to use a complex network structure to extract abstract features. Although achieving superior detection performance, the increase in network complexity means an increment of calculation costs. We have summarized the floating point operations (FLOPs) of schemes in previous literature to show the relationship between accuracy and network complexity. As shown in Table 4, schemes with higher FLOPs tend to have better detection performance while better detection performance does not necessarily mean higher FLOPs. This is a trade-off between network complexity and detection effect. We believe that a wise solution should achieve higher precision detection with lower network complexity. Under such

TABLE 4 Detection performance on FaceForensic++ datasets

Study	Method	Dataset	Performance	FLOPs
Bonettini et al. [31]	Ensemble of CNNs	FaceForensics++(c23)	0.9444 (AUC)	0.24
Nguyen et al. [34]	A capsule network	FaceForensics++ - Face2Face	93.11% (ACC)	>7.72
Zhao et al. [48]	Optical flow	FaceForensics++ - DeepFake	98.10% (ACC)	0.24
Cozzolino et al. [76]	Noiseprint + siamese network	FaceForensics++	92.14% (ACC)	-
Rossler et al. [7]	XceptionNet	FaceForensics++(raw)	99.26% (ACC)	8.42
		FaceForensics++(c23)	95.73% (ACC)	
		FaceForensics++(c40)	81.00% (ACC)	
Afchar et al. [36]	MesoInception-4	FaceForensics++(raw)	95.23% (ACC)	0.5
		FaceForensics++(c23)	83.10% (ACC)	
		FaceForensics++(c40)	70.47% (ACC)	
Sabir et al. [77]	CNN + GRU + STN	FaceForensics++ - DeepFake	96.9% (ACC)	14.4
		FaceForensics++ - Face2Face	94.35% (ACC)	
		FaceForensics++ - FaceSwap	96.3% (ACC)	
Li et al. [32]	Face X-ray + multitask learning	FaceForensics++ - DeepFake	0.9912 (AUC)	>3.99
		FaceForensics++ - FaceSwap	0.9909 (AUC)	
		FaceForensics++ - Face2Face	0.9931 (AUC)	
		FaceForensics++ - NeuralTexture	0.9927 (AUC)	
Ciftci et al. [63]	Biological signals	FaceForensics++ - DeepFake	93.75% (ACC)	-
		FaceForensics++ - FaceSwap	96.25% (ACC)	
		FaceForensics++ - Face2Face	95.25% (ACC)	
		FaceForensics++ - NeuralTexture	81.25% (ACC)	
Tarasiou et al. [43]	A lightweight architecture	FaceForensics - DeepFake (c23)	97.90% (ACC)	-
		FaceForensics - Face2Face (c23)	98.58% (ACC)	
		FaceForensics - FaceSwap (c23)	98.32% (ACC)	
		FaceForensics - DeepFake (c40)	92.40% (ACC)	
		FaceForensics - Face2Face (c40)	87.11% (ACC)	
		FaceForensics - FaceSwap (c40)	91.26% (ACC)	
Wu et al. [49]	SSTNet	FaceForensics++(c23)	98.57% (ACC)	>8.42
		FaceForensics++(c40)	90.11% (ACC)	
Li et al. [74]	Patch &pair CNN	Faceforensics(raw)	0.996 (AUC)	1.82
		Faceforensics(c23)	0.983 (AUC)	
		Faceforensics(c40)	0.931 (AUC)	
Masi et al. [50]	Two-branch recurrent network	Faceforensics++(frames, c23)	0.987 (AUC)	-
		Faceforensics++(videos, c23)	0.9912 (AUC)	
		Faceforensics++(frames, c40)	0.8659 (AUC)	
		Faceforensics++(videos, c40)	0.911 (AUC)	

circumstances, summarizing previous algorithms and exploring new research directions are required to promote more effective detection algorithms. In this section, we will talk about some concerns over current detection methods and envision important directions that should receive more attention.

5.1 | Concerns

In view of current research on face-manipulated video detection, we have summarized the following concerns, which need significant attention in future research.

Study	Method	Performance	FLOPs
Bonettini et al. [31]	Ensemble of CNNs	0.8813 (AUC)	>0.04
Montserrat et al. [47]	An automatic weighting mechanism	91.88% (ACC)	>9.9
Tarasious et al. [43]	A lightweight architecture	88.76% (ACC)	-
Li et al. [32]	Face X-ray + multitask learning	0.892 (AUC)	>3.99
Mittal et al. [78]	Emotions behind audio and visual content	0.892 (AUC)	-
Selim Seferbekov	EfficientNet + task specific data augmentations	0.42798 (LogLoss)	72.35×8
Vert VertWM vert/ vert	Ensemble of WSDAN-based networks	0.42842 (LogLoss)	18.83
NtechLab	Mixup + EfficientNet + 3D conv	0.43452 (LogLoss)	72.35×3

TABLE 5 Detection performance on DFDC datasets

TABLE 6 Detection performance on Celeb-DF datasets

Study	Method	Performance	FLOPs
Dang et al. [42]	Multitask learning + attention mechanism	0.712 (AUC)	4.59
Li et al. [32]	Face X-ray + multitask learning	0.8058 (AUC)	>3.99
Ciftci et al. [63]	Biological signals detection	91.50% (ACC)	-
Tarasious et al. [43]	A lightweight architecture	92.62% (ACC)	-
Hernandez-Ortega et al. [72]	DeepFakesON-Phys(convolutional attention network)	91.50% (ACC)	0.48
Wang et al. [39]	Monitoring neuron behaviours	0.668 (AUC)	-

5.1.1 | Generalization

Generalization is an important indicator to measure the performance of algorithms, which is often adopted to evaluate the performance of the algorithm on unknown datasets. The detection algorithms proposed are mostly based on supervised learning, which is prone to overfit on their own datasets. Related experiments performed in [32] have proved that the generalization performance of existing detection algorithms is still insufficient for cross-dataset detection tasks. Each subdataset of FaceForensic++ is used as a training set to train the Xception network, which is then evaluated on the other subdatasets. Experimental data shown in Table 7 demonstrate that the Xception network is fragile when encountering unknown data, even reaching a detection accuracy of only 49.13%. Practically, there are great differences in the selection of source videos and postprocessing of generated videos, resulting in that different data sets often imply distinct distributions. To our knowledge, some work has focused on improving the generalization of algorithm [32, 38, 79, 80]. However, due to special design, these algorithms have their own inherent flaws. For example, face X-ray [32] heavily relies on blending steps, causing that it cannot detect artefacts in entirely synthetic images. Therefore, generalization is still an urgent problem to be solved.

5.1.2 | Interpretability

Interpretability has been an inherent problem for algorithms based on neural networks. As a black-box model, the neural

network cannot provide human-understandable justifications for its output. However, the detection algorithm must be interpretable in practical forensic scenarios, otherwise convincing results cannot be obtained. There has been some related work on interpretability in other fields [81–83], while research on interpretability has not progressed in deepfake detection fields. The interpretability of deepfake detection approaches is still an important issue needed to be solved in the future.

5.1.3 | Time consumption

When applied in a practical scene, time consumption becomes a significantly important point. In the foreseeable future, deepfake detection algorithms will be widely used on streaming media platforms to reduce the negative impact of deepfake videos on social security. However, current detection algorithms are far from wide implementation in practical scenarios due to their high time consumption. In this survey, we performed a brief evaluation on time consumption of existing neural networks. Specifically, we randomly select 10 videos from corresponding dataset for each trained model. Each video selected has a length of about 300 frames. In this evaluation, we detect 64 frames for each video so as to achieve more accurate results. The final time consumption is calculated by only considering the inferring time of models. As shown in Table 8, the average detection time of 10 videos is about 70–80 s, which means that each video spends 7–8 s to detect. Considering videos are much longer than 300

frames in the practical scenarios, such time consumption is far from meeting the needs of massive video detection. In the current literature related to deepfake detection, detection accuracy is regarded as the only standard while rare researches pay attention to the time consumption of deepfake detection. In the future, more attention should be devoted to studying how to design an efficient and high-accuracy detection method.

5.1.4 | Robustness

Robustness is often applied to evaluate the performance of detection algorithms when encounter various degradations. Compared with original videos, compressed videos are more difficult to detect because it ignores a lot of image information for higher compression rate. As shown in Table 4, detection algorithms often indicate a decrease in performance when encounter low-quality videos compared with high-quality videos. In addition to compression operations, videos may also encounter operations such as image reshape and rotation. Under such circumstances, robustness becomes an important property that must be considered when designing detection algorithms. An effective way to improve robustness should be to add a noise layer to the detection network, so that multiple data degradation scenarios are considered. Improving the robustness of existing detection methods would perform a significant role in the future.

5.2 | Future works

To address problems existing in current detection algorithms, we also envision some research directions, which will advance future research on face-manipulated video detection.

TABLE 7 Cross-dataset evaluation on FaceForensic++ dataset

Training set	Test set AUC				
	DF	F2F	FS	NT	FF++
DF	99.38	75.05	49.13	80.39	76.34
F2F	87.56	99.53	65.23	65.9	79.55
FS	70.12	61.7	99.36	68.71	74.91
NT	93.09	84.82	47.98	99.5	83.42

TABLE 8 Time consumption evaluation on FaceForensic++ subdataset

Model	Time consumption (s)											
	RAW				C23				C40			
	DF	F2F	FS	NT	DF	F2F	FS	NT	DF	F2F	FS	NT
EfficientNetB0	96.71	68.26	95.57	85.78	94.18	81.06	73.65	77.11	80.46	86.94	60.97	77.96
ResNet50	80.62	66.54	68.43	59.84	105.40	83.23	75.46	57.48	84.12	90.87	65.24	57.84
ResNet101	78.72	65.38	86.19	79.33	104.27	81.28	73.68	77.38	82.39	87.84	63.55	77.63
												AVERAGE

5.2.1 | Triplet training

The toughest problem for deepfake detection tasks is that generalization performance is not sufficient to support the needs of practical scenarios due to the different distribution of datasets. Under such circumstances, it is difficult for detection models to learn the intrinsic difference between real and fake videos. To address this problem, triplet training strategy would be a possible solution for such tasks [28, 31]. Triplet training aims to minimize the distance between samples with the same category and maximize samples between features with different categories in the feature space. Especially, the triplet training strategy ensures that the distance between samples with different categories is larger than the distance with the same category. Therefore, the optimization goal of triplet training would attempt to exploit the intrinsic difference between real and fake videos, providing assistance in subsequent classification tasks. In the field of face liveness detection, triplet training has been applied for domain adaptation tasks [84], demonstrating the potential of the triplet training strategy in finding intrinsic differences between real and fake videos, even if the datasets have different distributions.

5.2.2 | Multitask learning

Multitask learning, performing multiple tasks simultaneously, is proved to improve prediction performance comparing with single-task learning. Performing both forgery location and deepfake detection at the same time is found to be effective to improve accuracy in deepfake detection tasks. Multitask learning allows the model to perform two tasks at the same time, considering losses caused by both tasks, and further improving the performance of the model. In [32, 43, 85], also prove that forgery location plays a vital role in the deepfake detection task. Therefore, multitask learning has great potential for further improvement of deepfake detection.

5.2.3 | Antiforensics

Antiforensic technology is developed due to defects existing in current forensic technology. In the field of deepfake detection, neural networks are widely used to distinguish forgery videos. However, due to inherent defects, neural networks cannot resist attacks of adversarial samples [86–88]. To this end, researchers need to design more robust algorithms that can

withstand possible attacks found in the laboratory to prevent such attacks in real-world scenarios. The development of antiforensics technology can predict possible attacks in advance and discover the weakness of existing algorithms, thereby improving existing algorithms.

6 | CONCLUSION

In recent years, deepfake technologies, which rely on deep learning, are developing at an unprecedented rate. Malicious face-manipulated videos generated by deepfake algorithms can be rapidly disseminated through the global pervasiveness of the Internet, threatening social stability and personal privacy. To this end, commercial companies and research groups worldwide are conducting relevant researches to reduce the negative impacts of deepfake videos on people. In this article, we first introduce the generation technology of deepfake videos, then analyse the existing detection technology, and finally discuss the future research direction. Existing problems of current detection algorithms and promising research are particularly emphasized in this review. Generalization and robustness are particularly emphasized in this review. We hope this article would be useful for researchers engaged in deepfake detection research and restrain the negative impact of deepfake videos.

ACKNOWLEDGEMENTS

This work is supported in part by the Jiangsu Basic Research Programs-Natural Science Foundation under grant numbers BK20181407, in part by the National Natural Science Foundation of China under grant numbers U1936118, 61672294, in part by Six peak talent project of Jiangsu Province (R2016L13), Qinglan Project of Jiangsu Province, and '333' project of Jiangsu Province, in part by the National Natural Science Foundation of China under grant numbers U1836208, 61702276, 61772283, 61602253, and 61601236, in part by National Key R&D Programme of China under grant 2018YFB1003205, in part by the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD) fund, in part by the Collaborative Innovation Centre of Atmospheric Environment and Equipment Technology (CICAEET) fund, China. Zhihua Xia is supported by BK21+ programme from the Ministry of Education of Korea.

CONFLICT OF INTEREST

None.

PERMISSION TO REPRODUCE MATERIALS FROM OTHER SOURCES

Figure 3,4 comes from Reference [26]. Thus, we have added relevant references.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in <https://github.com/ondyari/FaceForensics>, reference number [1].

ORCID

Peipeng Yu  <https://orcid.org/0000-0003-0056-4300>

REFERENCES

- Chesney, B., Citron, D.: Deep fakes: a looming challenge for privacy, democracy, and national security. *Calif. L. Rev.* 107, 1753 (2019)
- Delfino, R.: Pornographic Deepfakes—Revenge Porn's Next Tragic Act—The Case for Federal Criminalization. 887, 88 *Fordham L. Rev.* (2019). SSRN 3341593
- Dixon, H.B., Jr.: Deepfakes: More frightening than photoshop on steroids. *Judges J.* 58(3), 35–37 (2019)
- Feldstein, S.: How Artificial Intelligence Systems Could Threaten Democracy. *The Conversation* (2019)
- Korshunov, P., Marcel, S.: Deepfakes: a new threat to face recognition? Assessment and detection. *arXiv preprint arXiv:1812.08685*. (2018)
- Li, Y., Chang, M.-C., Lyu, S.: Inictu 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 (2018)
- Rossler, A., et al.: Faceforensics++: Learning to detect manipulated facial images. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1–11 (2019)
- Li, Y., et al.: Celeb-DF: a new dataset for deepfake forensics. *arXiv preprint arXiv:1909.12962*. (2019)
- Dolhansky, B., et al.: The deepfake detection challenge dataset. *arXiv preprint arXiv:2006.07397*. (2020)
- Jiang, L., et al.: Deeper Forensics-1.0: a large-scale dataset for real-world face forgery detection. *arXiv preprint arXiv:2001.03024*. (2020)
- Chesney, R., Citron, D.: Deepfakes and the new disinformation war: the coming age of post-truth geopolitics. *Foreign Aff.* 98, 147 (2019)
- Bitouk, D., et al.: Face swapping: automatically replacing faces in photographs. In: *ACM SIGGRAPH 2008 Papers*, pp. 1–8 (2008)
- Yuan, L., et al.: Face replacement with large-pose differences. In: *Proceedings of the 20th ACM International Conference on Multimedia*, pp. 1249–1250 (2012)
- Zhang, X., Song, J., Park, J.I.: The image blending method for face swapping. In: 2014 4th IEEE International Conference on Network Infrastructure and Digital Content, pp. 95–98. IEEE (2014)
- Korshunova, I., et al.: Fast face-swap using convolutional neural networks. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 3697–3705 (2017)
- Olsewski, K., et al.: Realistic dynamic facial textures from a single image using gans. 2017 IEEE International Conference on Computer Vision (ICCV) (2017)
- Dale, K., et al.: Video face replacement. *ACM* (2011)
- Faceswap-gan. (2018). <https://github.com/shaoanlu/faceswap-GAN>
- Keras-vggface: Vggface Implementation with Keras Framework. (2019). <https://github.com/rcmalli/keras-vggface>
- Petrov, I., et al.: DeepFaceLab: a simple, flexible and extensible face swapping framework. *arXiv preprint arXiv:2005.05535*. (2020)
- Li, L., et al.: Faceshifter: towards high fidelity and occlusion aware face swapping. *arXiv preprint arXiv:1912.13457*. (2019)
- Vlasic, D., et al.: Face transfer with multilinear models. In: *ACM SIGGRAPH 2006 Courses*, p. 24. (2006)
- Thies, J., et al.: Face2face: real-time face capture and reenactment of RGB videos. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2387–2395 (2016)
- Suwajanakorn, S., Seitz, S.M., Kemelmacher-Shlizerman, I.: Synthesizing Obama: learning lip sync from audio. *ACM Trans. Graph.* 36(4), 1–13 (2017)
- Fried, O., et al.: Text-based editing of talking-head video. *ACM Trans. Graph.* 38(4), 1–14 (2019)
- Kim, H., et al.: Deep video portraits. *ACM Trans. Graph.* 37(4), 1–14 (2018)
- Thies, J., Zollhöfer, M., Nießner, M.: Deferred neural rendering: Image synthesis using neural textures. *ACM Trans. Graph.* 38(4), 1–12 (2019)
- Zhou, P., et al.: Two-stream neural networks for tampered face detection. 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2017)

29. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1251–1258 (2017)
30. Dolhansky, B., et al.: The deepfake detection challenge (DFDC) preview dataset. *arXiv preprint arXiv:1910.08854*. (2019)
31. Bonettini, E.D.C., et al.: Video face manipulation detection through ensemble of CNNs. *arXiv preprint arXiv:2004.07676*. (2020)
32. Li, L., et al.: Face x-ray for more general face forgery detection. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle, WA, USA pp. 5000–5009. (2020)
33. Amerini, I., et al.: Deepfake video detection through optical flow based CNN. *Proceedings of the IEEE International Conference on Computer Vision Workshops* (2019)
34. Nguyen, H.H., Yamagishi, J., Echizen, I.: 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. IEEE (2019)
35. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*. (2014)
36. Afchar, D., et al.: Mesonet: A compact facial video forgery detection network. In: *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pp. 1–7. IEEE (2018)
37. Szegedy, C., et al.: Going deeper with convolutions. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9 (2015)
38. Li, Y., Lyu, S.: Exposing deepfake videos by detecting face warping artifacts. *arXiv preprint arXiv:1811.00656*. (2018)
39. Wang, R., et al.: Fakespotter: a simple yet robust baseline for spotting ai-synthesized fake faces. *International Joint Conference on Artificial Intelligence (IJCAI)* (2020)
40. Liu, Z., et al.: Global texture enhancement for fake face detection in the wild. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020)
41. Nguyen, H.H., et al.: Multi-task learning for detecting and segmenting manipulated facial images and videos. *arXiv preprint arXiv:1906.06876*. (2019)
42. Dang, H., et al.: On the detection of digital face manipulation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5781–5790 (2020)
43. Tarasiou, M., Zafeiriou, S.: Extracting Deep Local Features to Detect Manipulated Images of Human Faces. In: *2020 IEEE International Conference on Image Processing (ICIP)*, pp. 1821–1825. (2020)
44. Güera, D., Edward, J.: Delp: Deepfake video detection using recurrent neural networks. In: *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pp. 1–6. IEEE (2018)
45. Szegedy, C., et al.: Rethinking the inception architecture for computer vision. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826 (2016)
46. Sabir, E., et al.: Recurrent-convolution approach to deepFake detection-state-of-art results on FaceForensics++. *arXiv preprint arXiv:1905.00582* (2019)
47. Montserrat, D.M., et al.: Deepfakes Detection with Automatic Face Weighting. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2851–2859. (2020)
48. Zhao, Y., et al.: Capturing the persistence of facial expression features for deepfake video detection. In: *International Conference on Information and Communications Security*, pp. 630–645. Springer (2019)
49. Wu, X., et al.: SSTNet: Detecting manipulated faces through spatial, steganalysis and temporal features. In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2952–2956. IEEE (2020)
50. Masi, I., et al.: Two-Branch Recurrent Network for Isolating Deepfakes in Videos. In: *16th European Conference on Computer Vision ECCV 2020*, pp. 667–684 Springer, Cham (2020)
51. Ruff, L., et al.: Deep one-class classification. In: *Proceedings of Machine Learning Research*, vol. 80, pp. 4393–4402. PMLR, Stockholm (2018)
52. Yang, X., Li, Y., Lyu, S.: 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 (2019)
53. Lukas, J., Fridrich, J., Goljan, M.: Digital camera identification from sensor pattern noise. *IEEE Trans. Inf. Forensics Secur.* 1(2), 205–214 (2006)
54. Chen, M., et al.: Determining image origin and integrity using sensor noise. *IEEE Trans. Inf. Forensics Secur.* 3(1), 74–90 (2008)
55. Chierchia, G., et al.: A bayesian-MRF approach for PRNU-based image forgery detection. *IEEE Trans. Inf. Forensics Secur.* 9(4), 554–567 (2014)
56. Korus, P., Huang, J.: Multi-scale analysis strategies in PRNU-based tampering localization. *IEEE Trans. Inf. Forensics Secur.* 12(4), 809–824 (2016)
57. Koopman, M., Rodriguez, A.M., Geradts, Z.: Detection of deepfake video manipulation. In: *The 20th Irish Machine Vision and Image Processing Conference (IMVIP)*, pp. 133–136. (2018)
58. Frank, J., et al.: Leveraging frequency analysis for deep fake image recognition. *arXiv preprint arXiv:2003.08685*. (2020)
59. Cozzolino, D., Verdoliva, L.: Noiseprint: ACNN-based camera model fingerprint. *IEEE Trans. Inf. Forensics Secur.* 15, 144–159 (2019)
60. Zhang, K., et al.: Beyond a gaussian denoiser: Residual learning of deep CNN for image denoising. *IEEE Trans. Image Process.* 26(7), 3142–3155 (2017)
61. Huang, Y., et al.: Fakeretouch: Evading Deepfakes Detection via the Guidance of Deliberate Noise (2020). *arXiv preprint arXiv:2009.09213*
62. Chen, C., et al.: Camera trace erasing. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2950–2959 (2020)
63. Gıftci, U.A., Demir, I., Fakecatcher, L.Y.: Detection of synthetic portrait videos using biological signals. *IEEE Trans. Pattern Anal. Mach. Intell.* 1 (2020)
64. Donahue, J., et al.: Long-term recurrent convolutional networks for visual recognition and description. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2625–2634 (2015)
65. Feng, L., et al.: Motion-resistant remote imaging photoplethysmography based on the optical properties of skin. *IEEE Trans. Circ. Syst. Video Technol.* 25(5), 879–891 (2014)
66. Kumar, S., Prakash, A., Tucker, C.S.: Bounded kalman filter method for motion-robust, non-contact heart rate estimation. *Biomed. Optic. Express.* 9(2), 873–897 (2018)
67. Zhao, C., et al.: A novel framework for remote photoplethysmography pulse extraction on compressed videos. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1299–1308 (2018)
68. Chen, W., McDuff, D.: DeepPhys: Video-based physiological measurement using convolutional attention networks. In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 349–365. (2018)
69. Fernandes, S., et al.: Predicting heart rate variations of deepfake videos using neural ODE. In: *Proceedings of the IEEE International Conference on Computer Vision Workshops* (2019)
70. Chen, R.T.Q., et al.: Neural ordinary differential equations. In: *Advances in Neural Information Processing Systems*, pp. 6571–6583 (2018)
71. Qi, H., et al.: DeepRhythm: exposing deepfakes with attentional visual heartbeat rhythms. *arXiv preprint arXiv:2006.07634*, 2020
72. Hernandez-Ortega, J., et al.: DeepFakesON-phys: deepfakes detection based on heart rate estimation. *arXiv preprint arXiv:2010.00400*. (2020)
73. Korshunov, P., Marcel, S.: Vulnerability assessment and detection of deepfake videos. In: *The 12th IAPR International Conference on Biometrics (ICB)*, pp. 1–6 (2019)
74. Li, X., et al.: Fighting against deepfake: Patch&pair convolutional neural networks (PPCNN). In: *Companion Proceedings of the Web Conference 2020*, pp. 88–89 (2020)
75. Dufour, N., Gully, A.: Deepfakes Detection Dataset (2019)
76. Cozzolino, D., Poggi, G., Verdoliva, L.: Extracting camera-based fingerprints for video forensics. In: *Proceedings of the IEEE Conference*

- on Computer Vision and Pattern Recognition Workshops, pp. 130–137 (2019)
77. Sabir, E., et al.: Recurrent convolutional strategies for face manipulation detection in videos. *Interfaces*. 3 (2019)
 78. Mittal, T., et al.: Emotions dont lie: an audio-visual deepfake detection method using affective cues. In: *Proceedings of the 28th ACM International Conference on Multimedia, MM 20*, pp. 2823–2832. Association for Computing Machinery, New York (2020)
 79. Cozzolino, D., et al.: Forensicttransfer: Weakly-supervised domain adaptation for forgery detection. *arXiv preprint arXiv:1812.02510*. (2018)
 80. Du, M., et al.: Towards generalizable forgery detection with locality-aware autoencoder. *arXiv preprint arXiv:1909.05999*. (2019)
 81. Scott, L., Lee, S.: A Unified Approach to Interpreting Model Predictions, pp. 4768–4777 In: *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*, (2017)
 82. Samek, W., Wiegand, T., Muller, K.: Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models. *arXiv:Artificial Intelligence* (2017)
 83. Kermany, D.S., et al.: Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*. 172(5), 1122–1131 (2018)
 84. Jia, Y., et al.: Single-side domain generalization for face anti-spoofing. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8484–8493 (2020)
 85. Stehouwer, J., et al.: On the detection of digital face manipulation. *arXiv preprint arXiv:1910.01717*. (2019)
 86. Carlini, N., Farid, H.: Evading deepfake-image detectors with white-and black-box attacks. *arXiv preprint arXiv:2004.00622*. (2020)
 87. Gandhi, A., Jain, S.: Adversarial perturbations fool deepfake detectors. *arXiv preprint arXiv:2003.10596*. (2020)
 88. Neekhara, P., et al.: Adversarial deepfakes: Evaluating vulnerability of deepfake detectors to adversarial examples. *arXiv preprint arXiv:2002.12749*. (2020)

How to cite this article: Yu, P., et al.: A Survey on Deepfake Video Detection. *IET Biom*. 10(6), 607–624 (2021). <https://doi.org/10.1049/bme2.12031>