Multimedia Tools and Applications
https://doi.org/10.1007/s11042-022-13100-x

Emergence of deepfakes and video tampering detection approaches: A survey

Staffy Kingra[1] · Naveen Aggarwal[1] · Nirmal Kaur[1]

Received: 24 September 2020 / Revised: 17 February 2022 / Accepted: 4 April 2022

Abstract
Digital content, particularly the digital videos recorded at specific angle, though, provides a truthful picture of reality but the widespread proliferation of easy-to-use content editing softwares doubt about its authenticity. Recently, Artificial Intelligence (AI) based content altering mechanism, known as deepfake, became popular on social media platforms, wherein any person can be able to purport the behaviour of another person in a video who is actually not there. Depending on the type of manipulation performed, different types of deepfakes are described in this paper. Moreover, rely on digital content for trustworthy evidence as well as to avoid spread of misinformation, integrity and authenticity of digital content has-been of utmost concerns. This paper aims to present a survey of the state-of-art video integrity verification techniques with special emphasis on emerging deepfake video detection approaches. Seeing the advancement in creation of more realistic deepfake videos, this review facilitates the development of more generalized methods with a thorough discussion on different research trends in the wake of deepfake detection.

Keywords Video forensics · Video forgery detection · Deepfake detection · Facial manipulation detection · Audio deepfake detection

1 Introduction

Digital images, the idea of which was generated about a century ago, have acquired an important place in almost every field be it entertainment, education or human sentiments. With technological advancement, the idea of capturing scenes was extended to video recording. Besides this, image capturing and video recording devices became portable from

✉ Staffy Kingra
staffysk@gmail.com

Naveen Aggarwal
navagg@gmail.com

Nirmal Kaur
nirmaljul19@gmail.com

[1] University Institute of Engineering and Technology, Panjab University, Chandigarh, India

Published online: 05 August 2022

computers to cameras and later on mobile phones, thereby mitigating the time and place-related constraints to capture/record a visual scene. These captured or recorded events have been gaining importance to serve as an important piece of evidence in case of court proceedings, personal disputes, and other sensitive matters. Notably, digital videos serve to provide sufficient evidence as it contains the recording of a whole scenario.

With an increase in the usage of digital media, numerous techniques have also been developed to process or edit the video after its production. These post-production techniques were primarily designed to enhance the digital content with respect to its colour, contrast, and brightness. However, such video editing techniques were proliferated towards malicious manipulation that compromised the authenticity of video content.

Can we trust the digital content? For utilizing a digital video as an evidence, reliability and integrity are of utmost importance. Owing to the wide-spread usage of post-production techniques to alter digital content intentionally, integrity of content is hard to rely on. Such malicious alteration to digital video is known as video forgery/ tampering/ doctoring. Recently, in 2018, White House press secretary Sarah Sanders claimed an aggressive behaviour of CNN reporter Jim Acosta by sharing a video on twitter [60], wherein Acosta swiftly chop down on one side of an arm while Sarah tussled to grab microphone from Acosta. However, a video shared by Acosta's supporter justifies Acosta's arm movement only as a response to Sarah's tussle. The glimpse of the video is shown in Fig. 1. According to news, Sarah posted a doctored video in which Acosta's arm was appeared to move promptly by increasing frame rate of the original video sequence. In the light of such forgery cases, it is evident that a slight change in video content or frame rate can be harmful for the dignity of individual and society at large.

Recently, many other kinds of intentional manipulations are reported, where a source person in the video is swapped with another person (target), and intended to say things said by the source person. Famous politician Barack Obama have also been the victim of this type of forgery. Recently, a doctored video of Facebook CEO, Mark Zuckerberg, was uploaded on Instagram in which he was made to say things he never actually said [142]. Figure 2 shows a glimpse of AI (Artificial Intelligence) created deepfake video. As per the Deeptrace report [6], deepfake targeted industry mostly belongs to entertainment (62.7%) and fashion (21.7%). Famous politicians and business tycoons are the main victims of such deepfaked video, and this malpractice is escalating high day by day. In view of this, one can certainly imagine a world where anyone is made to say anything and do anything in digital videos; consequences of which can eventually diminish trust in news and digital media.

1.1 Video tampering: Analysis of different tampering techniques

Until recently, many mobile and desktop applications such as Lightworks, GIMP, Snapchat etc. have been designed to enhance the content of digital multimedia. Being user-friendly and inexpensive, such applications are widely being exploited by users to enhance and manipulate digital media. However, such easily accessible editing applications have raised serious integrity issues being faced by the research fraternity nowadays. It is pertinent that many video editing applications, which were primarily developed for entertainment, were later on utilized for malicious purposes. In the present scenario, it has been so easy to delude the eye of anyone with tampered videos or images. Advent of deep learning and easily available data has added more realism in the manipulated content.

Fig. 1 Glimpse of White House secretary's doctored video [60]

1.1.1 Video forgeries with respect to granularity levels

Depending on the type and amount of manipulation a user wants to perform, video tampering can be possible from fine to coarse grain level. Falsification performed using the content of different video sequences, either by pasting an object from one video to another or by combining the content of two video sequences, is referred to as an inter-video forgery. On the other hand, falsification performed within a single video sequence is termed as intra-video forgery; it is usually performed either by targeting different objects of the same frame or different frames of the same video sequence as described below.

1. **Inter-Frame Forgery:** Tampering performed with an intention of manipulating a sequence of frames by insertion, removal, or duplication. Frame insertion forgery can cause the occurrence of any unusual incident possible in the video that is not actually present. On the other hand, frame removal or frame replication can conceal an act;

FACEBOOK
ZUCKERBERG:
WE'RE INCREASING TRANSPARENCY ON ADS
ANNOUNCES NEW MEASURES TO "PROTECT ELECTIONS"
CBSN

Fig. 2 Glimpse of Mark Zuckerberg's doctored video [142]

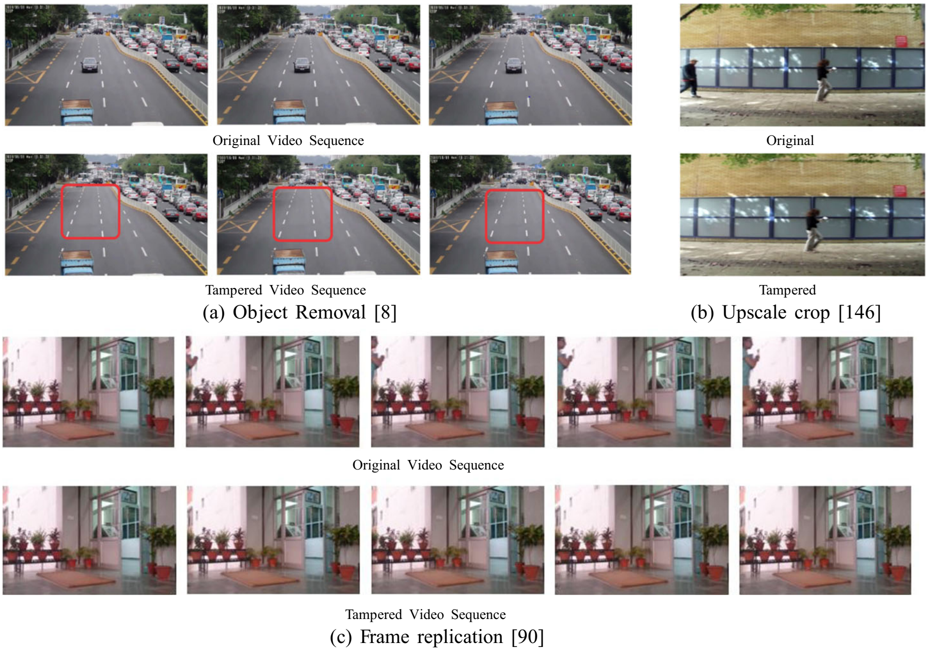
these are relatively easy to perform but difficult to detect if done carefully. One of the inter-frame forgery i.e. frame replication is illustrated in Fig. 3c.

2. **Intra-Frame Forgery:** Tampering performed only on a particular frame is referred to as intra-frame forgery. It can be performed either on the whole frame, on different objects of those frames, or on mere pixels. Pixel level forgery [91] works on the intensity information of the frame, usually by inducing noise to blur the existence of any object. In contrast, block and frame-level forgery works on insertion or removal of a specific object or region.

1.1.2 Video Forgeries with respect to the nature of manipulation

At each granule of the video sequence be it pixel, frame or the whole video sequence, content can be manipulated in various forms. Video forgery, on the basis of the type of manipulation performed, is explained here.

1. **Object-based manipulation:** Performed at intra-frame level,it deals with insertion/removal of objects to/from specific locations, and is termed as object insertion/removal tampering respectively. To conceal the mismatched area after removal of an object, video inpainting [204] is usually performed. As illustrated in Fig. 3a, an object (car) shown in the original sequence is removed in the tampered version. This object-based tampering can also be used to perform copy-paste forgery in a video sequence by cloning an object from one place and pasting it to another.



(a) Object Removal [8]

(b) Upscale crop [146]

(c) Frame replication [90]

Fig. 3 Demonstration of different forgeries (a) and (b) shows intra-frame tampering while (c) shows inter-frame tampering

2. **Upscale Crop:** It is performed at intra-frame level, where the whole frame is first enlarged and then cropped from the extremities to get the original size [159]. This forgery eventually conceals the occurrence of an object present at the boundary. A still from a video sequence is shown in Fig. 3b, by which a man present near the left boundary of the frame is removed to create a tampered version.
3. **Frame rate up-conversion (FRUC):** A forgery performed by a mere increase/decrease in frame rate can totally alter the way of interpreting any scenario by reducing/increasing the visual focus on any object or event. This forgery, however, is very easy to perform but a mere tricky to detect. It is performed by inserting interpolated frames periodically into the video sequence. These interpolated frames are generated using frame replication or averaging or Motion Compensated Frame Interpolation [202].
4. **Real-time Frame Forgery:** Recently, a smart attack was investigated that is performed against smart VSS (Video Surveillance System) at the time of recording a video [128]. However, VSS was designed to capture a scene only when triggered; this trigger can be a change in light intensity or a slight movement of object etc. This VSS design has been made target by forgers who trigger the injection of false frames (false frame injection attack) or replay some previous frames (replication attack) in real-time on the detection of a specific activity.

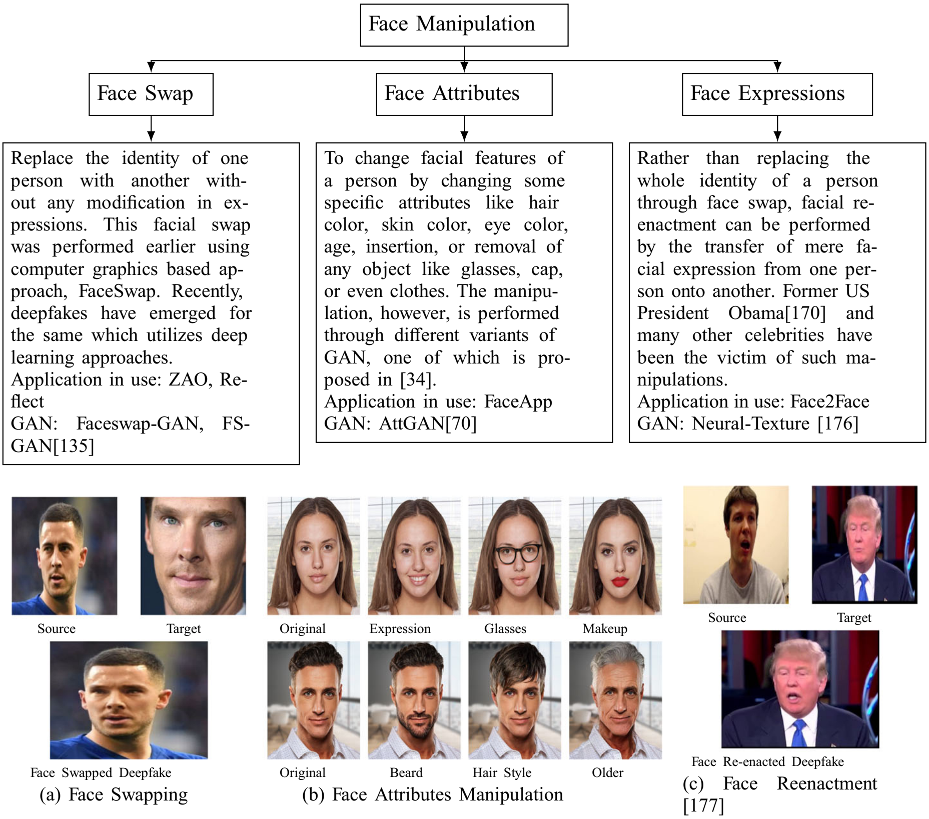
1.2 Deepfake and its creation

Until recently, many specialized applications have been built to perform manipulation on the video content. Emergence of AI into this field has greatly influenced the society. An AI created tampering technique, called Deepfake, provides the ability to map a person's facial or voice features (source) with some other person's (target) in a video clip with an intention of making the target person saying or doing things that were actually said or done by source. The landmark of such animations was set in 1997 using a Video Rewrite program [18]. Deepfakes became popular with the release of GAN[1] which has the ability to generate anything from previously trained data. Facial synthesis came into picture in 2017 when popular celebrity faces like Gal Gadot, Emma Watson, Hilary Duff, and Jennifer Lawrence were used in porn content [40]. Afterwards, many celebrities have been the victim of AI-based deepfaking technology, first by syncing the lip movement corresponding to a particular audio clip and then by making person a puppet. In view of this, various types of deepfakes created, so far, are described here:

1. **Face Manipulation:** With the emergence of deep learning, quality of performing face manipulations in videos have also been improved. It can transfer facial expressions(facial features) from one person to another on one hand, and change facial attributes of the target face on the other as demonstrated in Fig. 4. This strategy was then utilized to create a user-friendly application called FakeApp[2]. A new application 'DeepNude' was developed in mid-2019 by which any person can be made unclothed with GAN technique. One of the limitations of GAN networks for face synthesis was the requirement of a large amount of training data, which is eliminated with the proposal of MocoGAN [15].

[1] GAN: Generative Adversarial Network

[2] http://www.fakeapp.com/



Face Manipulation

Face Swap

Replace the identity of one person with another without any modification in expressions. This facial swap was performed earlier using computer graphics based approach, FaceSwap. Recently, deepfakes have emerged for the same which utilizes deep learning approaches.
Application in use: ZAO, Reflect
GAN: Faceswap-GAN, FS-GAN[135]

Face Attributes

To change facial features of a person by changing some specific attributes like hair color, skin color, eye color, age, insertion, or removal of any object like glasses, cap, or even clothes. The manipulation, however, is performed through different variants of GAN, one of which is proposed in [34].
Application in use: FaceApp
GAN: AttGAN[70]

Face Expressions

Rather than replacing the whole identity of a person through face swap, facial reenactment can be performed by the transfer of mere facial expression from one person onto another. Former US President Obama[170] and many other celebrities have been the victim of such manipulations.
Application in use: Face2Face
GAN: Neural-Texture [176]

(a) Face Swapping (b) Face Attributes Manipulation (c) Face Reenactment [177]

Fig. 4 Deepfakes using Face Manipulation

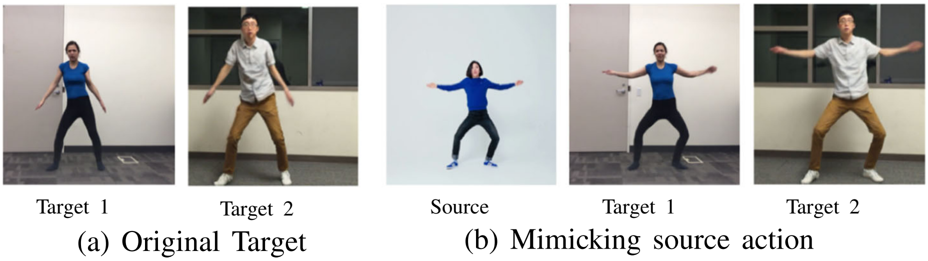
The principle behind face manipulation deepfake is to separately train two autoencoders in parallel, one with the faces of source person and another with the target one. Encoders of both share their weights to blend the features of source and target while decoders are kept separate to reconstitute specific characteristics of a person's face. It makes two autoencoders encoded using the same feature set with illumination information, position information, and facial features, and keeping separate the morphological and contextual information. Deepfake is, then, generated by encoding the source person using a shared encoder and decoding it with the decoder of target person.

2. **Lip-sync Deepfake:** Besides facial synthesis, many techniques have been designed to make the mouth region consistent with a particular audio recording. For instance, a video of Barack Obama altered by an actor and director Jordan Peele went viral where Obama was made to say things that were actually said by Jordan Peele [183]. A similar technique was proposed in [169], which requires a large amount of training data of a target person. The concept of lip-syncing is not new; many techniques became popular in the past which synchronize the mouth of movie actors w.r.t. audio sequence by constructing a 3D face model [59]. Since it is difficult to map a 1-D audio signal to 3-D mouth and lip movements of a person, the conventional computer graphics results were not much convincing. Deep learning technique proposed in [169] produced real-

istic videos, though not actually real, by training a generator with an ample amount of images of target person. Another deep learning based approach ‘Neural Voice Puppetry’ [174] was also made popular to render the target face corresponding to a specific audio with an advantage of using less amount of training data in comparison to [169].

3. **Puppet-master:** In addition to lip-sync, head movement, eye movements, and other gestures of a person have also been made realistic in puppet-master. This new technique learn the full body features of source person and overlay these features to a target person by which a target person is made to imitate movements of source person [54, 191]. Apart from these, Face-GAN [22] was proposed to transfer dance movements of the source person to target one as demonstrated in Fig. 5. In the figure, two target persons are shown that are imitating the movements of source person. Such fake images were created in earlier times using a non-deep learning approach [29, 71, 200], which have been now improvised using deep learning mechanisms [12, 181]. Although this technique doesn’t seem perfect until now, deepfakes created using these are as convincing to create a risk of being misused.
4. **Audio Deepfake:** Beyond continuously revamping image synthesis approaches, another AI-oriented illusion of the fake person speaking in a voice similar to real person has launched. Till now, forgers could be able to generate a fake email or web page purporting it to be from some other person. Through AI-based audio synthesis, a cybercrime was reported in late 2019, wherein a forger scammed a company for millions of dollars by imitating voice of CEO of the same company [78], which started a new trend of audio deepfakes. Two audio synthesis approaches namely Voice Cloning (VC) and Text-to-Speech (TTS) exists in literature. In the former, one’s voice is swapped with another while the later is used to convert text to audio. WaveNet [137], Sprocket [90] and Tacotron [195] are the prominent models utilized for VC and TTS synthesis.

Many companies incorporated AI into their technologies to perform editing of images and videos in different ways. For instance, a company named pinscreen developed an app that can create different digital avatars of any person by manipulating audio or/and video content [104]. Moreover, Adobe’s VoCo and Lyrebird has made the creation of fake audio recordings possible. Though developed for entertainment, users have started using these in malicious ways. It has put the authenticity and validation of such video sequences in uncertainty. With the emergence of such deepfakes, ways to detect these are also getting attention. Since with every new detection technique, forgers try an anti-forensic approach and make an improved deepfake, the arms race between generating fakes and fighting fakes seem to be on an endless road. To maintain the authenticity of video evidence, US government



(a) Original Target (b) Mimicking source action

Fig. 5 Stills from Puppetry Deepfake [22]

revealed that digital content can be introduced as a law of evidence in the court[3,4] as long as its authenticity, accuracy, reliability, and trustworthiness can be established with respect to US Federal Rules of Evidence [117].

Research in multimedia forensics [161, 163, 184] has been carrying on for more than 15 years by academic researchers, IT companies, and even by major funding organizations like DARPA[5]. Despite continuous and joint efforts, the advent of AI has changed the game by introducing deepfakes, more realistic faked content, detection of which needs more efficient and timely solutions. To maintain the credibility of such videos, numerous techniques have been developed and many others are being developed with improvements in previous ones. A brief overview of the creation and detection of deepfake videos is provided in [132]. Recently, authors in [179] and [182] provided a good survey of deepfake detection techniques containing a thorough analysis of facial synthesis detection approaches.Although a sufficient literature of deepfakes is provided in [123] and authors discussed in-depth about different generation architectures. However, the paper does not contain detailed analysis of deepfake detection approaches. These state-of-art deepfake focused surveys did not pay attention to another video forgeries prevailing from the beginning. Also, among different deepfakes, only facial synthesis detection approaches were discussed.

In this paper, accompanying facial synthesis, a brief overview of other kind of deepfakes proposed for deceiving the person's identity by faking lip movement (lip-sync), body movement (puppet-master), and voice (audio deepfakes) is provided. Since the advent of deepfake forgery has not put an end to other traditional tampering techniques, this survey also presents a review of different approaches developed for detecting previous-era video forgeries. Section 2 provides a detailed overview of state-of-art techniques developed for detecting different types of video forgeries with special emphasis on deepfake detection techniques in Section 3. A concise and important outline of discussed techniques is also provided in tabular form in respective sections. Section 5 then concludes the survey with a thorough discussion of some different research directions that can be followed in the future for more efficient detection of deepfake videos.

2 Conventional video forgery detection methods

It is pertinent that drastic increase in the availability of video editing software has made the detection of manipulated videos a prerequisite in multimedia domain. These video forgeries often leave some footprints which, if analyzed accurately, can aid in distinguishing tampered videos from benign ones. Video forgery detection can be performed using active or passive approach. Active approaches embed a unique watermark or signature into the video sequence, during or after video recording, to authenticate video content. However, it is hard to acquire specialized software for embedding which also degrades the video quality. On the other hand, the passive counterpart for video forgery detection relies on unusual intrinsic artifacts (static and temporal), which are supposed to be introduced into the video sequence due to some post-production manipulation. Passive approaches for video forgery detection are, therefore, considered better than active approaches, and are in the current focus of research.

[3] United States v Beeler, 62 F Supp. 2d 136 (July 1, 1999, United States District Court, D. Maine).

[4] Dolan v State of Florida, 743 S. 2d 544 (July 21, 1999, Court of Appeal of Florida, Fourth District).

[5] Defense Advanced Research Project Agency

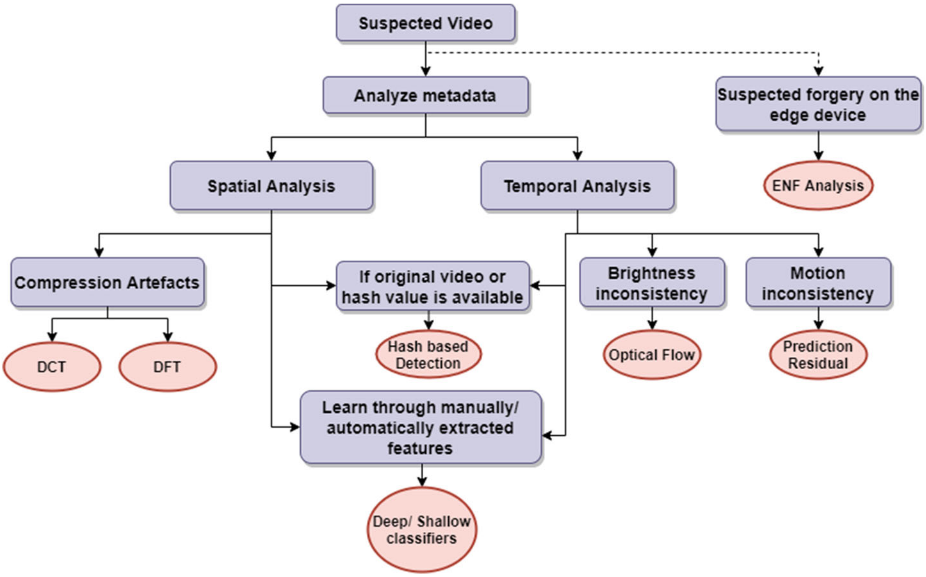


Fig. 6 Visual Media Tampering Detection Approaches

Numerous researchers proposed, and are continuously proposing different types of passive video forgery detection techniques to detect a variety of video forgeries. This section provides a review of most of the techniques proposed in literature; some of which can be extended further for better detection. Figure 6 demonstrates the use of different forgery detection approaches under different scenarios. It was observed from the literature that there is a lack of standardized datasets pertaining to conventional video forgeries. Thereby, most of these approaches were evaluated on videos generated by the respective authors themselves. However, some of the available datasets are provided in Table 1 Moreover, evaluation metrics used by these state-of-art techniques are provided in Table 2.

2.1 Forgery detection in different transform domain

Video is considered as a particular sequence of frames which are correlated with respect to time. Processing on a frame is similar to how an image is processed. Some features of an image are best analyzed in time domain and others in frequency domain. Thereby, researchers tried to analyze different artifacts from varying domains such as Discrete Cosine

Table 1 Datasets available for conventional video forgery detection

Dataset	Source	Tampering type	Videos	Format
SULFA [145]	Canon SX220, Nikon S3000, Fujifilm S2800HD	Copy-move	150	MOV & AVI
REWIND [144]	SULFA	Copy-move	10	MOV & AVI
VTD [7]	Internet	Splicing, copy-move, frame-swapping	33	AVI
GRIP [43]	Internet	Copy-move	15	AVI

Table 2 Evaluation metrics used by state-of-art forgery detection approaches

Metric	Description
Accuracy	Measure of correctly predicted data instances with respect to total data instances.
Precision	Measure of correctly predicted positive data with respect to all positive predicted data.
Recall	Measure of correctly predicted positive data with respect to all positive data.
F1-Score	Provides a weighted average of precision and recall.
AUC	Measure of degree of separability between two classes of particular model.

Transform (DCT), Discrete Fourier Transform (DFT), and Discrete Wavelet Transform (DWT) to detect forgery as discussed in this section.

Any type of video manipulation is performed by extracting frames, manipulating these frames separately, and then re-saving manipulated frames as a tampered video. Re-saving of digital content leads to double compression. So, in the early days of development of video forgery detection approaches, it was believed that analysis of double compression ensures the presence of manipulation in video. As double compression disturbs the distribution of DCT coefficients, most of the authors have analyzed the same in different ways to detect forgery as shown in Table 3. The very first method [192] to detect frame insertion/removal forgery in a video sequence analyzed double compression in MPEG video sequences by detecting periodicity in DCT coefficient distribution of I-frames, and motion error generated by P-frames. Although the author did not formulate any quantitative results in the paper but claimed that the technique's performance boosts if number of frames deleted/inserted are multiples of 3. Same group of researchers, then, extended the technique in [193] where they also detected composite videos. Double compression was detected by analyzing Gaussian distribution of DCT coefficients. Unlike the previous technique, it detected forgery even if different quantization scales are used in first and second compression. The technique developed in [165], rather than analyzing peak and periodicity, examined convex pattern in the histogram generated through DCT coefficients of each macro-block. Howbeit, selection of a different threshold for different set of videos was found somewhat inconvenient.

Later on, it was stated in [122] that even a non-tampered video can have double compression artifacts due to mere transmission, uploading, or downloading. So, detection of double compression can't be the only key to ensure forgery. This brought up the need to develop techniques that are independent of double compression. The same author [122] developed an approach that utilized statistics of first digits of quantized DCT coefficients and extract 63-D features, which were fed to the SVM classifier. Focusing on same goal, differences in DCT energies were computed as MCEA (Motion Compensated Edge Artifact) and plotted in FFT spectrum. Presence of spikes in the resultant FFT spectrum revealed the presence of frame insertion/removal forgery. DFT domain was also utilized to detect MCFI based Frame-rate Up-conversion (FRUC) forgery [107], by analyzing irregularity in the noise level of frames. Interpolated frames must have lower noise levels relative to other frames, which can be easily analyzed using high pass filtering.

2.2 Hash-based video forgery detection

Many video forgery detection techniques utilized the concept of hash function wherein, hash value obtained from specific features of a suspected video was analyzed with respect

Table 3 DCT or DFT based forgery detection (A: Accuracy, TPR: True Positive Rate, TNR: True Negative Rate)

Technique	Dataset	Results	Performance
Periodicity in I frames and prediction error of P frames [192]	Two MPEG-1 encoded videos	No quantitative analysis	Dependant on number of frames tampered and re-quantization performed.
Gaussian distribution of DCT coefficients of I-frames [193]	Three MPEG-2 encoded videos	A: 99.4% (q>1.7) A: 41.2% (1.3<q<1.7) A: 2.5% (q<1.3)	Performance depends on quantization values.
Convex pattern in histogram of DCT coefficients [165]	100 MPEG-2 encoded videos	TPR: 98-100% TNR: 93%	Performance depends on output bit rate. Inefficient for slow videos.
SVM classification using 63-D features [122]	12 videos	A: >73%	Detected number of re-compressions. Performance degrades with increase in compression.
MCEA [50]	4 MPEG-2 video	No quantitative analysis	Efficient for multiples of sub-GOP deletion
Noise estimation in Fourier domain [107]	YUV video sequences	A: 100% (in most cases)	Proved efficient for different MCFI methods Low complexity

to some threshold or original video. Hash value of a video is considered efficient for manipulation detection if it is robust against content-preserving operations [138]. This hashing can be applied either on each frame individually, termed as frame-based hashing, or on the whole video, termed as spatio-temporal hashing.

Some of the frame-based hashing techniques developed for video tampering detection were proposed in [44, 102, 103]. Centroid of Gradient Orientation (CGO) utilized in [102] was robust against compression and noise but sensitive for innocuous geometrical transformations. Overcoming this, hash value was computed using radial projections [44] and affine covariance [103]. The former one was claimed robust against rotation while the later was robust against all geometric or non-geometric transformations. In addition, image hashing methods [139, 207] can also be utilized for frame-level hashing. Frame-based hashing method does not consider temporal nature of the video, which makes hash value less sensitive to deliberately performed temporal de-synchronization (Table 4).

Spatio-temporal hashing was found relatively better which focuses on the temporal nature of a video too. Based on this, one of the researcher computed hash value by applying any image hashing algorithm on TIRI (Temporally Informative Representative Images) images [116]. TIRI image is just a representation of the whole video sequence obtained by the weighted average of luminance values of successive frames. This method was found robust against noise degradation and contrast enhancement. Another work proposed in [106] utilized PARAFAC (Parallel Factor Analysis) to derive a hash algorithm, wherein 2-dimensions of the video represent image content while 1-dimension represent temporal content. This technique was found more robust against content preserving operations like

Table 4 Hash-based video forgery detection (TPR: True Positive Rate, FRR: False Rejection Rate, FAR: False Acceptance Rate)

Technique	Dataset	Results	Performance
Hashing on TIRI images [116]	Video Traces [155]	TPR: 99.2%	Robust against noise degradation and contrast enhancement. Not robust to rotation.
Hashing using multilinear subspace projection [106]	Youtube videos	Average hash deviation: 0.2404 (normalized)	No quantitative analysis.
Hashing using DWT [151]	14 test videos	FRR: 1.8% FAR: 1%	Not robust against geometric operations like rotation.
Multiple level hashing using 3D-DWT and PCET [26]	Original: 100 Tampered: 100	Avg. Precision: 72% Avg. Recall: 97.2%	Robust against various inoffensive manipulations. Coarse-to-fine localization.

compression, contrast manipulation, blurring, frame rotation, cropping, scaling or change in frame rate etc. Later on, DWT [151] was utilized to compute the hash value based on the fact that video coding based on 3D-DWT exhibits inherent spatio-temporal scalability feature [1]. To enhance the robustness of hash value with respect to different content preserving operations, median thresholding was applied.

Not to mention, robustness against mild non-malicious operations and sensitiveness against illegal manipulations in a particular video are difficult to be fulfilled simultaneously [149]. To balance this trade-off, a combination of frame-level and spatio-temporal based hash was proposed in [26] wherein hash value was computed at multiple granularity levels. At frame sequence level, 3D-DWT was applied to get the lowest frequency sub-band LLL, and PCET moments were extracted to generate hash for that particular frame sequence. Here, 3D-DWT captures the spatio-temporal property of video while PCET extracts scaling and rotation invariant features. Subsequently, at block level, PCET moments of each block were computed to generate hash values of each annular and angular block. Pixel level hash, however, corresponds to object location detected using a saliency map.

2.3 Metadata based video forgery detection

Most of the techniques proposed in literature observed anomalies dependant on visual content. With advancement in technology, visual artifacts in tampered videos have been greatly lessened which induced the need for a different approach to analyze disruption in video content. One of the alternatives pursued in [64] examined metadata of the respective video using multimedia stream descriptors as shown in Table 5. However, ensemble of Random forest and SVM classifier trained with these descriptors provided best performance.

2.4 Optical flow based video forgery detection

Any variation in the location of object among subsequent frames create velocity vector corresponding to each pixel. These vectors formulate an optical flow by depicting the

Table 5 Metadata based video forgery detection (AUC: Area Under Curve)

Technique	Dataset	Results	Performance
Training a binary classifier using multimedia stream descriptors [64]	MFC [61]	F1-score: 0.917 AUC: 0.984	Effective if footprints in metadata have not been sanitized.

movement of object with respect to an observer/scene. By assuming the brightness constancy constraint of pixels in successive frames, Horn-Shunck [73] and Lucas-Kanade [115] proposed two different algorithms for optical flow computation. Different variants of optical flow have been used in different computer vision applications like object tracking [11], action recognition [136], face recognition [118], motion estimation [171], etc.

In video forensics domain also, optical flow has been used in different ways to detect forgery in video sequences as demonstrated in Table 6. One of the technique analyzed optical flow from adjacent frame-pair [23] while the other trained SVM using extracted optical flow [194]. Besides, authors in [187] utilized Gaussian model to detect discontinuities in the optical flow which tends to unveil one discontinuity point in case of frame deletion and two in case of frame insertion and duplication. Beyond detection of frame insertion, frame removal, and frame duplication, optical flow has also been used for detection of copy-move forgery [16] in video sequences. The technique [16] extracted particular ROI (Region of Interest) from a frame, that is supposed to contain large motion. After computing optical flow of each ROI, the author computed the ratio of optical flow of each frame with its previous and following frame; this ratio tends to have periodic pattern for the original sequence only.

Table 6 Optical flow based video forgery detection (P: Precision, A: Accuracy, TPR: True Positive Rate)

Technique	Dataset	Results	Performance
Window and frame-pair based analysis [23]	KTH database	P: 98% (frame insertion) P: 89% (frame deletion)	Fast execution. Inefficient for mild tampering.
SVM classifier [194]	598 videos from 5 video datasets	Accuracy: 96.75%	Differentiate frame insertion and frame deletion forgery.
Gaussian model [187]	MPEG-2 videos (TRECVID) [180]	A: 90%	Limited dataset. Complex approach.
Autocorrelation between adjacent frame's ROI [16]	REWIND [144] SULFA [145]	TPR: 85.7%	Inaccurate for large motion videos.
GMM (Gaussian Mixture Model) distribution [154]	SULFA [145] wisdom Weizmann	A: 83.3% (Detection) A: 71.8% (Localization)	Tested on both complex and conventional inpainted videos.
Markov chain statistics [47]	GIF and HD720p [41]	A: 99.12%	Efficient for highly quantized videos.

On the other hand, technique proposed in [154] computed optical flow of each frame with respect to its m successive frames, providing m optical flow matrices for a specific frame. RMSE of Chi-square value obtained from histograms of these matrices was considered as distinguishing artifact. Recently, optical flow has also been utilized for the detection of FRUC forgery in video sequence [47]. The technique computed optical-flow based prediction residual by finding a matched pair for each pixel in adjacent frames. These feature vectors computed using four different optical flow algorithms were considered as states of Markov chain and fed to an ensemble classifier.

2.5 Prediction residual based video forgery detection

Prediction residual is the measure of difference between original frame and version of same predicted from its previous frame. Large value of prediction residual raises inconsistency among subsequent frames, which may be due to deletion or insertion of frames in-between. Hence, prediction residual has largely been seen to detect inter-frame video forgery. Most of the methods proposed using prediction residual have been explained thoroughly in [88]; a handful of techniques described here are demonstrated in Table 7.

In [89], prediction residuals were utilized along with the optical flow for inter-frame forgery detection. The principle behind this technique was that higher the variation between prediction residuals or optical flows of the subsequent frames, greater would be the chance of tampering performed. Similar concept was utilized in [160] where optical flow and prediction residuals were used to analyze irregularities caused by frame insertion/removal and frame replication respectively. Later on, deblocking filter [74] of H.264/AVC was utilized to detect the insertion/removal of a particular shot into/from a video sequence. The technique computed blocking strength to analyze variation in prediction residual.

2.6 ENF (Electrical Network Frequency) based video forgery detection

ENF signals are embedded with video or audio recorded by any system connected to a power grid, such as surveillance system, which provides variation in the frequency of power supply. This surveillance feed can be monitored by extracting the time-varying nature of ENF signal using Fourier Transform spectrum. Until recently, many algorithms have been proposed to detect forgery in already recorded and stored video files. Real-time frame forgery, as discussed in Section 1.1.2, can't afford late detection. ENF provides the capability of instant detection of malicious manipulation, in surveillance camera, performed on-site. Since ENF signals extracted from the audio recordings were found more reliable and efficient [58] as compared to video, the technique proposed in [127] focused on ENF signals from audio

Table 7 Prediction Residual based video forgery detection (AA: Average Accuracy)

Technique	Dataset	Results	Performance
Optical flow gradient and prediction residual gradient [89]	CCTV and Mobile recorded videos	AA: 83%	Less efficient for highly illuminated videos.
Optical flow and prediction residual [160]	SULFA [145] TRACES	AA: 98.6%	Worked well in the presence of noise too.
Deblocking filter [74]	14 YUV CIF videos	AA: 72%	Diverse experimentation.

Table 8 ENF based video forgery detection

Technique	Dataset	Results	Performance
ENF estimation of audio using FFT [127]	Self created	No quantitative analysis	ENF of video signal can also be utilized for reliability.

recordings only. This technique extracted ENF signal using Fourier Transform directly from the power supply; extraction was performed at multiple locations to maintain feasibility. Raspberry pi model was utilized as an edge device with a sound card inserted in it to record power and audio signals at the same time for comparison (Table 8).

2.7 Deep learning based video forgery detection

Classification of real and fake videos through machine learning classifier requires manually extracted features for training. These handcrafted features usually become obsolete over time on the advancement of tampered media. Deep learning classifier, however, extract features automatically, provided a large amount of data is available for training. Such networks learn fine to coarse detail from data with each successive layer. It has set its foot in almost every impossible field of computer vision like image classification [97, 98], speech recognition [205], text-to-speech synthesis [125], rainfall estimation [25], autonomous vehicles [177], and many more. Considering its importance, deep neural networks has been utilized in video forgery detection too [114, 203], as demonstrated in Table 9. The technique proposed in [114], C2FDCNN[6], detected duplication of frames in video sequence using two variants of ResNet (I3D and Siamese). I3D network analyzed video sequences at coarse level by computing distance matrix between overlapping subsequences while the Siamese network computed frame-to-frame distance matrix of selected frames. Duplication was detected by comparing the distance with threshold. Meanwhile, another technique [203] was proposed for detection of computer-generated videos by modifying pre-trained models of GoogleNet and ResNet[7]. In this paper, Q4 and cobalt filters were utilized based on DCT coefficients and re-quantization errors. Extracted features were fed to networks for training.

2.8 Discussion

Since digital videos are able to provide a valuable evidence in various critical matters, confirming its authenticity is of utmost importance. Over the years, video content authentication techniques has come a long way from traditional manual feature analysis to deep learning based automatic feature extraction and from recorded video analysis to on-site analysis. Digital forensics has gained a lot over the past few years, but the road is still ahead of us. This section provided a brief of conventional forgery detection approaches. One of the major shortcoming found in these approaches is the lack of multi-faceted tampering detector. Forensic researchers should focus on designing a comprehensive forensic system that can detect tampering regardless of its type. Moreover, video authentication techniques should also consider the role of audio modality for decision-making. On the other hand, quality of tampered media is increasing day by day. Along with conventional forgeries, deepfake forgeries are more prevailing now-a-days which can hinder one's identity in digital

[6] Coarse-to-Fine Deep Convolutional Neural Network

[7] Residual Network

Table 9 Deep learning based video forgery detection (AUC: Area Under Curve)

Technique	Dataset	Results	Performance
Coarse-to-Fine CNN [114]	MFC [61] Self collected	AUC: 98.84% (MFC) AUC: 82.75% (author's)	Localize tampering. Focused on frame duplication.
GoogleNet and ResNet [203]	MFC [61] InVID [140]	AUC: 80.6% (within dataset) AUC: 64.73% (Cross dataset)	Frame level annotation of frame should be there.

video. Increasing risk of deepfaking in real-life scenarios has shifted the attention of forensic researchers towards deepfake detection. State-of-art techniques proposed for deepfake detection are surveyed in next section.

3 State of art in the detection of deepfake video forgery

A wide range of strategies emerged with the inception of AI, which simulates human behavior and thought process by making the AI machine learn through ample amount of data. Some of the first generation's AI-based techniques include Amazon's Alexa, Apple's Siri, and Pandora's automated music recommendation service. These AI-based mechanisms are continuously improving their performance and behavior by learning from new inputs. Additionally, with the emergence of Google's Tensorflow, an open-source AI tool for machine learning and image processing, many unthinkable ventures have been made possible. Deepfakes are also a result of AI advancement and defending of deepfakes can also be possibly done by the same. However, this section puts a light on most of the state-of-art deepfake detection techniques and datasets built for the same.

3.1 Benchmark datasets for deepfake detection

Various datasets are released in the said domain to evaluate the robustness and generalization of proposed techniques on continuously ameliorating deepfakes, progression of which is shown in Fig. 7. To lessen visual manipulation artifacts in deepfake videos, each dataset utilized an improved version of generation algorithm relative to previous one as detailed in Table 10. Face manipulation on *FaceForensics* dataset performed using computer graphic techniques, was utilized to evaluate some deepfake detection techniques. Seeing the increasing need for deepfake detection techniques, *UADFV, Deepfake-TIMI,* and *FaceForensics++* datasets were released. These novice datasets suffer from certain limitations being either small, exhibit visual inconsistencies, or of low quality. Besides this, lack of enough subjects and content in datasets lead to overfitting of detection model.

Afterwards, Google company contributed to solve this significant problem prevailing in the society and released a dataset, named*DFD*[8] (DeepFake Detection Dataset) in collaboration with Jigsaw. *Celeb-DF* and *DeeperForensics* further improved the quality, quantity, and realism of generated deepfaked content. As a result of high demand and risk of these deepfakes being misused, Facebook along with other reputed organizations also commenced a

[8] This dataset is available as a part of FaceForensics.

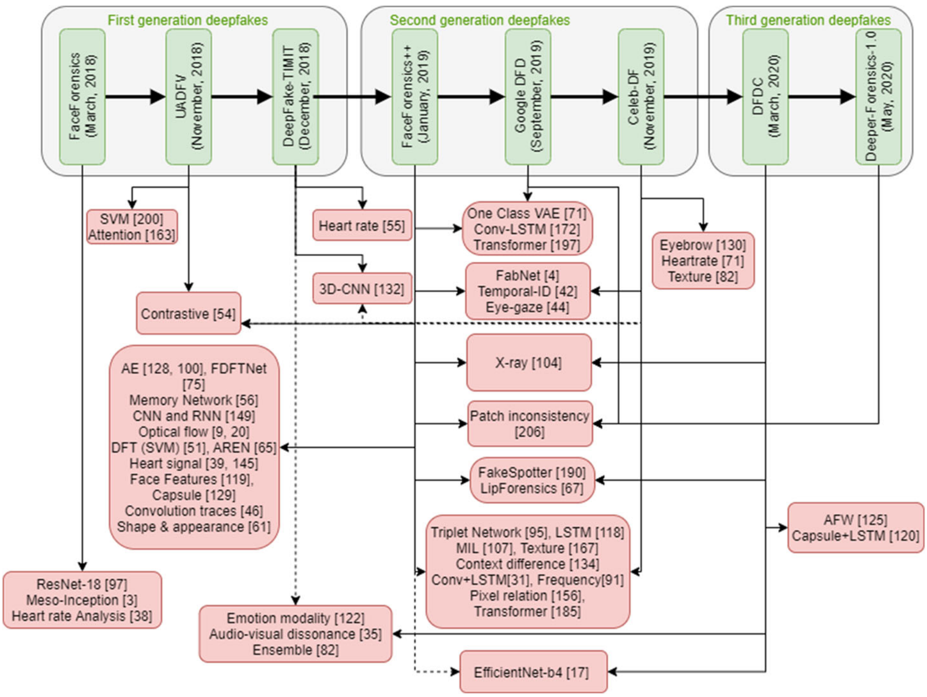

First generation deepfakes
FaceForensics (March, 2018)
UADFV (November, 2018)
DeepFake-TIMIT (December, 2018)
Second generation deepfakes
FaceForensics++ (January, 2019)
Google DFD (September, 2019)
Celeb-DF (November, 2019)
Third generation deepfakes
DFDC (March, 2020)
Deeper-Forensics-1.0 (May, 2020)
SVM [200] Attention [163]
Heart rate [55]
One Class VAE [71] Conv-LSTM [172] Transformer [197]
Eyebrow [130] Heartrate [71] Texture [82]
3D-CNN [132]
FabNet [4] Temporal-ID [42] Eye-gaze [44]
Contrastive [54]
AE [128, 100], FDFTNet [75] Memory Network [56] CNN and RNN [149] Optical flow [9, 20] DFT (SVM) [51], AREN [65] Heart signal [39, 145] Face Features [119], Capsule [129] Convolution traces [46] Shape & appearance [61]
X-ray [104]
Patch inconsistency [206]
FakeSpotter [190] LipForensics [67]
AFW [125] Capsule+LSTM [120]
Triplet Network [95], LSTM [118] MIL [107], Texture [167] Context difference [134] Conv+LSTM[31], Frequency[91] Pixel relation [156], Transformer [185]
ResNet-18 [97] Meso-Inception [3] Heart rate Analysis [38]
Emotion modality [122] Audio-visual dissonance [35] Ensemble [82]
EfficientNet-b4 [17]

Fig. 7 Evolution of Deepfake Datasets and Detection Techniques (Green boxes represent databases and Red boxes show techniques that utilized the specified databases)

challenge called DFDC[9] and released a dataset for the same. Moreover, to introduce diversity, different perturbations, such as noise, compression, blur etc., are performed on videos contained in current generation deepfake datasets. Best performed techniques, submitted in DFDC, for deepfake detection are reported in [48].

3.2 Facial manipulation detection

Facial manipulation is usually performed either by swapping a face with another person, manipulating facial attributes or expressions of the target person as discussed in Section 1.2. To detect facial manipulation, different ways proposed by researchers are classified here with respect to the technique utilized ranging from handcrafted features training to smart contract embedding as demonstrated in Tables 11, 12, 13 and 14.

3.2.1 Handcrafted features for deepfake detection

In mid-2018, during first generation of deepfake videos, it was realized that DNN[10] based face mapping, even at fine level, could not generate every facial feature much realistic. Considering the fact that a normal person blink 17 times per minute[11] and eyes remain close

[9] DeepFake Detection Challenge

[10] Deep Neural Network

[11] https://www.ncbi.nlm.nih.gov/pubmed/9399231.

Table 10 Deepfake video datasets for evaluating deepfake detection techniques (LQ: Low Quality, HQ: High Quality)

Dataset	Tampering Performed	Technique Utilized	Forged Videos	Real Videos	Original Video Dataset	Total perturbations	Utilized Subjects	Best performed Technique
UADFV	Deepfake	FakeApp	49	49	CEW	-	49	[55]
Deepfake-TIMIT	Face Swap	Faceswap-GAN	320 (LQ) 320 (HQ)	320	VidTIMIT	-	32	[84]
Face Forensics++	Face Swap	Self-designed (CG)	1000	1000	Youtube (news anchor, face videos)	2	Not disclosed	[3, 31, 55, 99]
	Face reenactment	Face2Face (CG)	1000					
	Face reenactment	Neural Texture Model (GAN)	1000					
	Face Swap	DeepFake Model (Autoencoder)	1000					
Google DFD	DeepFake	Improved DF Model	3068	363	Paid actors	-	28	[206]
Celeb-DF	Face Swap	Self-designed autoencoder architecture	5639	590	Youtube	-	59	[31, 157]
Deeper Forensics-1.0	Many-to-many face swap	DeepFake VAE	10000	50000	Paid actors	35	100	[206]
DFDC	Face Swap	DFAE (128L) DFAE(256L) MM/NN face swap Neural talking head FSGAN STYLE-GAN	100000	19154	Paid actors	19	960	[84, 126]
	Audio Swap	TTS						

(Best performed techniques are elaborated later in this paper)

Table 11 Handcrafted features for Deepfake Detection (AUC: Area Under Curve, A: Accuracy)

Technique	Dataset	xResults	Performance
CNN with LRCN [109]	CEW [162] and self created	AUC: 99%	Future deepfakes counter the effect of this signal.
CNN and LSTM [65]	HOHA [100] (Real videos) Randomly collected (Deepfake videos)	A: >97%	Different attacks must be analyzed to increase robustness.
CNN (ResNet and DenseNet) and RNN [150]	FaceForensics++ [148]	A: 96.9% (DF) A: 94.35% (F2F) A: 96.3% (FS)	Tested on front face videos only. Limited videos; a multi-recurrence model can't be applied.
SVM with facial landmarks [201]	Created dataset (UADFV) MFC [61]	AUROC: 86.65%	Not suitable if deepfakes are advanced.
MLP and Logistic Regression trained using different set of features [120]	CelebA [113] ProGAN [79] Glow [86]	AUC: 86.6%	Dataset dependant
CNN trained using optical flows [8]	FaceForensics++ [148]	A: 81.61% (VGG16) A: 75.46% (ResNet50)	Less accurate.
Train SVM, LR and K-means using Fourier transform features [53]	CelebA [113] DFD [51]	A: 100% (High resolution images) A: 91% (low resolution videos)	Lesser amount of data required. Effective only for high resolution images/videos.
Frequency Analysis [92]	FF++ [148] Celeb-DF [110]	A: 85.24% A: 66.50%	Visualized Activation Maps.
ResNet-18 models on local face regions [99]	FaceForensics [147]	A: 96.75%	More efficient for uncompressed videos.
Face X-ray [105]	Training data: FaceForensics++ [148] DFD [51] DFDC [49]	AUC:> 98% (known data) AUC:>80% (unknown data)	Dependant on blending artifacts. Effective for high resolution videos.
Analysis of convolutional traces [63]	Authentic data: CELEBA Tampered data: STYLEGAN [80] (9999) STYLEGAN2 [81] (3000) Author created data using: STAR-GAN [34] (5648) ATTGAN [70] (6005) GDWCT [32] (3369)	Maximum A: 90.22% (5-NN)	Dependant on kernel size Not evaluated on benchmark datasets.
Detected traces of up-convolution [52]	Faces-HQ [a] CelebA [113] FaceForensics++ [148]	A: 100% A: 100% A: 90%	Possible to conceal this generator's limitation in future deepfakes.

Table 11 (continued)

Technique	Dataset	xResults	Performance
Optical flow based CNN [20]	FF++ [148]	A: 97%	Inefficient cross forgery performance.
Eye-gaze tracking [45]	FF++ [148] Celeb-DF [110] DFD [51]	A: 89.79% A: 88.35% A: 80%	Not generalized.
Eyebrow matching [131]	Celeb-DF [110]	AUC: 87.9%	May not be effective for long-term.
LBP+HRNet [83]	Celeb-DF [110] DFDC-P [49]	A: 86% A: 76.8%	Used fine-tuning for generalizability.

[a] Download from: https://cutt.ly/6enDLYG

for a period of 0.1-0.4 seconds per blink[12], irregular blinking was earlier considered as a deepfake detector [109]. This detection was performed using VGG-16 variant of CNN[13] at frame level along with an LRCN[14] at temporal level. Here, CNN detected eye-state (open or closed) in individual frames while LRCN compared the same with neighboring frames. A similar perspective was followed in [65] to exploit features at frame level and scene level using CNN and LSTM[15] respectively. CNN was trained to exploit any inconsistency that prevails along the boundary of swapped face with respect to its surroundings. In contrast, LSTM exploited inconsistency, say illumination variation and flickering, between faces in subsequent frames. Features of each frame extracted from CNN were concatenated to provide an input to LSTM for temporal comparison.

Most of the deepfake methods proposed earlier were able to detect video forgery only if the person is facing towards the camera, making the technique inefficient in case of different face alignments. To cope with it, different alignments of face were analyzed to detect deepfake forgery [150]. These alignment features were extracted by applying CNN variants (ResNet and DenseNet), which were fed into RNN for temporal analysis. Moreover, this model extracted features at the micro, meso, and macro level which was made possible through DenseNet. Another technique [201] extracted facial landmarks from video to analyze the 3-D head pose of a person. Then, SVM classifier was trained with these landmark features extracted from real and deepfake videos. Meanwhile, facial reenactment detection technique [120] based on three different classifiers was proposed. These classifiers were trained using three different sets of facial features; color difference in left and right eye, a combination of inconsistent eye and teeth details, and a combination of irregular face and nose border. However, a combination of eye-teeth and face-nose features provided efficient results.

This goal was further pursued in [8] by analyzing optical flow of concerned video sequence to exploit inter-frame dissimilarities. Since deepfaking is performed on each face/frame individually, deepfaked face among subsequent frames may have subtle motion discontinuity. Optical flow was supposed to uncover this unusual lip, eye, and face movements in deepfaked video. Here, CNN based PWC-Net followed by Flow-CNN (semi-trainable) was utilized for the purpose. Researchers made use of a transfer learning

[12] http://bionumbers.hms.harvard.edu/bionumber.aspx?id=100706&ver=0.

[13] CNN: Convolutional Neural Network

[14] LRCN: Long-Term Recurrent CNN, a combination of CNN and LSTM

[15] LSTM: Long Short Term Memory

Table 12 Deep neural networks for Deepfake Detection (AUC: Area Under Curve, A: Accuracy, DR: Detection Rate)

Technique	Dataset	Results	Performance
Meso-4 and MesoInception-4 [2]	FaceForensics [147] and self-created	DR: 98% (Deepfake) DR: 95% (Face2Face)	Analysis at meso-scopic level
Capsule Network [130]	Face Swap [2], Face reenactment [147], replay attaked [30] videos	A: 97%	Analyzed in the presence of noise. Tested on many attacks. Frame level analysis.
Attention layer with deep network [164]	UADFV Celeb-DF	A: 99.64%	Diverse variety of deepfaked videos.
Fake Detection Fine-tuning Network [76]	Celc ebA [113] PGGAN [79] Deepfake [147, 148]	AUC: 97.02%	Claimed to detect unseen data.
Hierarchical Memory Network [57]	FaceForensics [147] FaceForensics++ [148] FFW [85]	A: 98.75% (Known data) A: 85.33% (Unknown data)	Provide good results on compressed data. Able to generalize to unseen data.
FakeSpotter [189]	FaceForensics++ [148] Celeb-DF [110] DFDC [49]	A: 98.5% A: 66.8% A: 68.2%	Performance of layers dependant on the datasets and Face Recongnition architecture.
Triplet Network using XceptionNet, RNN and 3-D convolution [96]	Celeb-DF [113] FaceForensics++ [148]	AUC: 92.9%	Still to be tested on current generation deepfakes.
EfficientNet-b4 (end-to-end and siamese training) [17]	Images from: FaceForensics++ [148] DFDC [48]	Using ensemble: A: 94.4% A: 87.82%	Similar frames from videos for training may lead to overfitting. Large amount of training data but not good performance.
Adaptive Residual Extraction Network (ARENnet) [66]	Fake faces generated using: PGGAN [79] StyleGAN [80] Glow [86] Face2Face [147] StarGAN [34]	A: 93.99% (Average)	Less generalizable. Not much efficient for FaceForensics++ dataset. Need to evaluate on current generation deepfakes.
Contrastive Learning [55]	FF++ [148] UADFV [201] Celeb-DF [113]	A: 99.9%	Need triplet pair.
Convolutional Vision Transformer [198]	FF++ [148] UADFV [201] DFD [51] DFDC [48]	A: 71.8% A: 93.75% A: 91% A: 91.5%	Good accuracy on DFDC

technique on some part of network and fine-tuned the rest. Later on, other technique [53] utilized frequency domain analysis wherein 2-D representation of FFT spectrum of images was transformed to 1-D. These reduced feature vectors were inputted to Logistic Regression, SVM, and K- means classifier individually. Following similar approach, authors of

Table 12 (continued)

Technique	Dataset	Results	Performance
Multi-scale transformer [186]	FF++ [148] Celeb-DF [113]	A: 99.5% A: 95.5%	Need to evaluate on advanced deepfakes.
FAb-Net with ResNet (Behavioural features) VGG-16 (Appearance features) [3]	WLDR FaceForensics++ [148] DFD [51] DFDC-P [49] Celeb-DF [110]	A: 94.14%	Low performance on advanced deepfakes.
ID-matching [42]	FF++ [148] Celeb-DF [113] DFD [51]	A: 85% A: 65% A: 80%	Less efficient on face re-enactment deepfake.
Automatic Face Weighting [126]	DFDC [48]	A: 92.61%	Inefficient for low-quality videos.
Capsule Network and LSTM [121]	DFDC [48]	A: 83.42% AUC: 91.15%	Tested on self-created deepfakes also. Unable to outperform XceptionNet.
2-branch Bi-LSTM network [119]	FaceForensics++ [148] Celeb-DF [113]	A: 93.18% A: 73.41%	Inefficient on advanced deepfakes.
Conv-LSTM Residual Network [173]	FF++ [148] DFD [51]	A: > 96%	Need to evaluate on advanced deepfakes.
LipForensics [67]	Celeb-DF [113] FF++ [148] DFDC [48] Deeper-Forensics [77]	A: 87.7%	Perform only if mouth is altered.
3D-CNN [133]	FF++ [148] DF-TIMIT [94]	A: 99.33% A: 99.2%	Need to evaluate on current generation deepfakes.
Y-shaped auto-encoder [129]	FaceForensics [147] FaceForensics++ [148]	A: 83.71% (Classification) A: 93.01% (Localization)	Can deal with unseen attacks Require testing on compressed videos
FakeLocator [75]	CelebA [113] FFHQ [80]	A: 99.85% (Maximum)	Generalization applicability for short term.
Audio-Visual dissonance based model [36]	DFDC [48] DeepFake-TIMIT [94]	AUC: 91.5% AUC: 96.5%	Only audio-exhibiting video sequences can be evaluated. May provide false detection for non-synchronizing audio and video.
Pixel Region Relation Network [157]	FF++ [148] Celeb-DF [113] DFDC-P [49]	A: 90.18% 99.80% 97.78%	Very few epochs.
Emotion modality from audio-visual cues [124]	DFDC [48] DeepFake-TIMIT [94]	AUC: 84.4% AUC: 95.6%	Need to train real video and its deepfake in pair. Emotion features are person-specific.

Table 12 (continued)

Technique	Dataset	Results	Performance
Face and context discrepancies [135]	FF++ [148] Celeb-DF [113]	A: 75.0% A: 66.0%	Need to evaluate on advanced deepfakes.
Invariant Texture Learning [168]	FF++ [148] Celeb-DF [113]	AUC: 86.4%	Considered only face replacement dataset.
Patch-wise inconsistency [206]	FF++ [148] DFD [51] Celeb-DF [113] DFDC [48] Deeper-Forensics [77]	AUC: 99.79% AUC: 99.07% AUC: 94.165% AUC: 67.53% AUC: 99.41%	Not much efficient for advanced deepfakes.
Inconsistency between 3D facial shape and facial appearance [62]	FF++ [148]	A: 87.3%	Need to evaluate on advanced deepfakes.
One Class VAE [82]	FaceForensics++ [148] DFD [51]	A: 88.28% (Average)	Best performance on Neural Textured data. Not evaluated on current generation deepfakes.
Multiple-instance learning [108]	DFDC [48] Celeb [113] FFPMS (from FF++)	A: 85.11% A: 98.84% A: 84.28%	Robust to compression. Focus on partially attacked videos.
Transfer Learning based AutoEncoders [101]	FF++ [148] Deepfakes in the wild	A: 99.8% A: 89.49%	Fine-tuning for generalizability seems impractical.
Ensemble [84]	Deepfake-TIMIT [94] DFDC [48]	A: 99.68% A: 96.50%	Not much efficient for high resolution deepfake videos.

Table 13 Biological Signals for Deepfake Detection (AUC: Area Under Curve, A: Accuracy)

Technique	Dataset	Results	Performance
Analysis of heart rate [39]	Deepfakes, FaceForensics [147]	A: 77.33% (Deepfake) A: 82.55% (FaceForensics)	Robust against compression and illumination. Complex technique. Manual feature extraction.
NeuralODE trained using heart rate [56]	Deepfake-TIMI [94] Deepfake created from: COHFACE VidTIMIT	Loss: 3.11%	New dataset generated No quantitave analysis.
Convolutional Attention Network [72]	Celeb-DF [113] DFDC-P [49]	A: 98.7% A: 94.4%	Inefficient results in externally illuminated videos.
Spatio-temporal attention [146]	FF++ [148] DFDC-P [49]	A: 98% A: 64.1%	Inefficient for current generation deepfakes.

Table 14 Blockchaining smart contract for Deepfake Detection (AUC: Area Under Curve, A: Accuracy)

Technique	Dataset	Results	Performance
History tracking using block chaining [68]	Deepfakes	Not implemented yet	Can ensure integrity, non-repudiation and authorization. Need to automate this technique.

[92] also extract frequency domain features of facial regions using Global DCT (GDCT) and train a 3-layer CNN based classifier on extracted data.

Later on, an approach [99] proposed for facial re-enactment detection analyzed local artifacts from each face region that were supposed to be suppressed in case of highly compressed faces. It was performed by training five ResNet-18 (Residual Network) models in parallel; four with different face regions, and one with the whole face. ResNet-18 model was chosen because of its robustness against highly compressed videos. In addition, a new loss function was designed to avoid bias towards any of the five models, which eventually ameliorated the performance of proposed model. Performance of most of deepfake detection models depends on data with which it is trained and thereby can't generalize on unseen data. Considering the generalization objective, researchers started focusing on the methodology being adopted to generate deepfakes. It was realized that most of the deepfake generators don't efficiently blend the target face onto the source frame. Thereby, the boundary of blended part of an image was exploited, using Face X-ray [105], which was supposed to be present only in fake videos. However, the utilized model HRNet required both manipulated image and its real version for generating a mask. As claimed more generalizable, this completely different approach reported good performance on both seen and unseen data.

By analyzing different artifacts for deepfaked media detection, it can't be denied that fake media synthesizers left its footprints in synthesized content. Thereby, local correlation of pixels in AI-generated image is said to be dependant on operations performed by layers in deepfake generator. These pixels get correlated when the same kernel (of varying sizes) gets convolved with different regions of image. Based on this belief, this pixel's relationship was captured using EM algorithm in [63]. Different classifiers (K-NN, SVM, LDA) were trained with feature vector obtained using EM algorithm to differentiate authentic images from forged ones. Instead, other researchers [52] discovered that every kind of synthesizer either GAN or VAE[16] utilizes up-sampling to increase resolution. Use of convolution up-sampling causes distortions in high-frequency distribution of synthesized data which can be analyzed by DFT spectrum of the suspected face. They also proposed to use large kernels for generator convolutions and also spectral regularization to make generation procedure much stable.

Recently, some researchers train CNN with normalized optical flows extracted from individual frames [20]. Some state-of-art approaches focused on specific part of face such as eye-gaze [45] and eyebrow [131] to determine the presence of deepfake. To track eye-gaze [45], authors extracted eye landmarks using OpenFace [14] and performed visual, geometric, temporal, spectral and metric analysis. Extracted features were fed into 3-layer dense network for training. On the other hand, for eyebrow recognition [131], four deep models were employed i.e. LightCNN, Resnet, DenseNet and SqueezeNet. Since deepfaking procedure tends to disrupt texture consistency, LBP (Local Binary Pattern) based histogram

[16] VAE: Variational AutoEncoder

combined with HRNet extracted features were utilzed for deepfake detection [83]. Here, Capsule based network performed classification.

Contrary to said detectors, authors of [190] claimed easy detection of CNN generated images, provided pre-processing, post-processing, and augmentations of training data is carefully performed. From high-pass filtered frequency spectra of each image, they analyzed that spectra of deepfaked images share a common periodic pattern. But this artifact keeps on curbing for continuously evolving GAN synthesizers and can't even be visualized for current deepfakes. Thereby, it would not be feasible to depend on visual artifacts or generator specific artifacts for deepfake detection.

3.2.2 Deep neural networks for Deepfake Detection

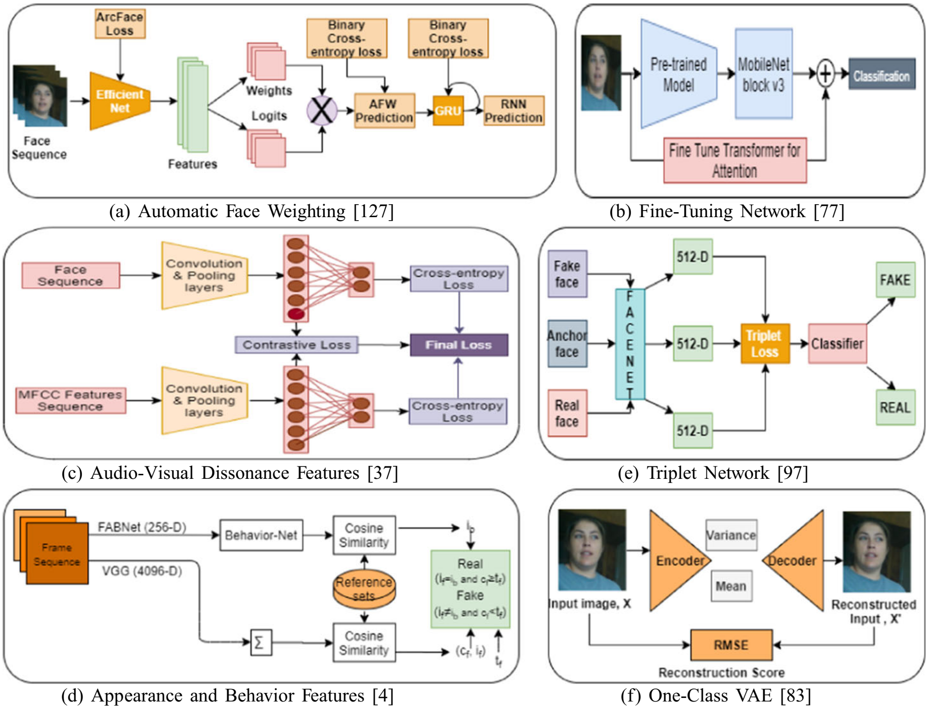
Since handcrafted features often become obsolete over-time, especially in the case of rapidly evolving quality of deepfakes, letting these trained through a deep network for deepfake detection is of no use. Deep learning researchers developed different variants of deep networks that analyze distinguishing features automatically. Some of the best performed models are demonstrated in Fig. 8. Meso-4 and MesoInception-4, proposed in [2], analyzed mesoscopic properties of individual frames for deepfake forgery detection using different sized kernels. Here, Meso-4 network utilized four convolution layers with max-pooling and two fully connected layers with a dropout of 0.5. MesoInception-4 model, on the other hand, utilized the inception module in place of four convolution layers of Meso-4.

Meanwhile, Capsule network [130] was utilized for the detection of swapped face in video sequence. Due to their hierarchical nature, capsule networks are able to learn orientation and relationship between different parts of an image. This network was built using three primary capsules and two output capsules to differentiate the real image from fake one. Here, primary capsules were trained using features extracted by VGG-19 network. This technique was also found good to detect replay attack, face reenactment and computer-generated content collected from different datasets. To further improve deepfake detection performance, attention-based layer was utilized in a variant of DNN [164] to emphasize most informative facial regions. In addition, a new dataset, DFFD (Diverse Fake Face Dataset), was also introduced by authors after collecting diverse variety of content from FFHQ [80], CelebA [113], and FaceForensics++ [148]. Some deepfake videos were created by manipulating facial attributes using FaceApp and StarGAN [33] and by manipulating the entire face using PGGAN[17] [79] and StyleGAN [80]. Using XceptionNet [35] as a backend to attention layer, the network provided a best accuracy of 99.64% on the collected dataset.

Detection of GAN-generated fake images and videos became quite more challenging with the release of few-shot learning technique [167], which can generate more realistic fake images using small amount of data. In view of this, a robust neural network- Fake Detection Fine-tuning Network (FDFtNet) [76] was proposed that make use of Fine-Tune Transformer (FTT) for feature extraction. In spite of good performance with lesser data, the technique proposed in [164] was found comparatively better in terms of accuracy and dataset quality.

On the contrary, social perception and cognition of human brain inspired some researchers to utilize memory networks to analyze visual cues stored in neural memories to reason the face and predict its future embeddings. Neuro-scientific researchers claimed the occurrence of uncanny valley effect in the human brain, on observing faces that lack

[17] Progressive Growing GAN



(a) Automatic Face Weighting [127]

(b) Fine-Tuning Network [77]

(c) Audio-Visual Dissonance Features [37]

(e) Triplet Network [97]

(d) Appearance and Behavior Features [4]

(f) One-Class VAE [83]

Fig. 8 Best performed models for Deepfake Detection

natural emotions and social presence [196]. Considering this, a Hierarchical Memory Network (HMN) [57] was designed combined with a multi-task learning approach to make the detector more infallible relative to the human brain. The technique provided relatively good results on highly compressed videos and images.

Afterward, FakeSpotter [189], assigned the task of spotting deepfakes onto the network itself and monitored neuron activation's behavior of any Face Recognition system. Since each layer plays its own role in learning input representations, activated neurons from each layer were tracked and fed into a shallow network. State of activated neurons was captured by feeding input images to VGG-Face and ResNet50 networks. Along with benchmark datasets, performance was also evaluated on fake faces created using the latest GAN's [81, 111] but was not found much efficient. Considering the performance of XcpetionNet, RNN, and 3D-Convolution, a triplet network was proposed [96] for deepfake detection. Triplet loss function, as described in (1), utilized for the triplet combination network ensures that embeddings of the fake face and pristine face remain well separated. In (1), p represents predicted output for anchor image (a), positive sample (p) and negative sample (n) with m being a hyperparameter. Due to its efficient detection on highly compressed videos, the technique has been claimed fruitful for authenticating media shared on internet too. Despite good results, the technique needs to be tested on more advanced deepfakes containing less visual artifacts.

$$L = max(|p(a) - p(p)|^2 - |p(a) - p(n)|^2 + m, 0) \tag{1}$$

Focusing on same goal, a group of developers proposed a modified version of EfficientNet-b4 [172] by employing an attention module [17]. Attention mechanism drove the model towards explainable AI by focusing on relevant parts of the frame. In addition, both end-to-end and siamese training strategies were adopted for training a model, resulting in four variants of the proposed model. EfficientNet-b4 (basic) and EfficientNet-b4Att (including attention layer) were trained using end-to-end approach while EfficientNet-b4ST and EfficientNet-b4AttST were trained using siamese strategy. To highlight tampering traces in fake images, ARENnet (Adaptive Residuals Extraction Network) [66] was proposed that employed AREN as a pre-processing block for CNN based detector. AREN was designed to predict image residuals that subtract the original image from the predicted feature map of image which eventually force CNN to learn features from manipulation traces. The technique was evaluated on diverse variety of fake face images[18].

Afterwards, researchers in [55] leveraged contrastive learning of modified XceptionNet architecture differentiate deepfaked images from pristine ones. Training utilized triplet loss function to minimize anchor-positive difference and maximize anchor-negative difference. Leveraging deep learning benefits, one of the approach [198] utilized transformer based model with stack of convolution layers as feature extractor. Features extracted from face images were fed into transformer, which consists of encoder, for classification. Since transformers are able to capture long-term information, some other scientists also proposed a Multi-modal Multi-scale Transformer (M2TR) [186] to detect local inconsistency amongst different patches. To improve robustness of proposed model, M2TR takes fusion of frequency and RGB features.

Spatio-Temporal Analysis While detecting manipulation in video sequences, analysis of temporal behavior along with spatial one always pays well. Based on this, an appearance and behavioral-biometrics based model [3] was utilized to capture facial expressions and head movements of a person in the video along with static facial features. Facial Attributes-Net (FAb-Net) [197] extracted facial movement and expression's features, that were made identity-specific by training ResNet-101 architecture. On the other hand, appearance features were captured using VGG-16 [141]. The video was classified real if appearance and behavioral features output similar identities and fake otherwise. Another identity specific deepfake detection architecture was proposed in [42] which train Temporal-ID network using real facial biometrics. This network was then utilized to extract 128-D embeddings of real images or 3DMM generated images, which would be compared with previously recorded data to identify deepfaking.

Another approach [126] leveraged EfficientNet-b5 [172] (initialized with ImageNet weights) and bi-directional GRU[19] for spatial and temporal analysis. The model also employed an AFW (Automatic Face Weighting) layer to provide weightage to features extracted by EfficientNet with respect to their relevance. Here, researchers employed three different loss functions; i) angular margin loss for EfficientNet to increase inter-class difference and reduce the intra-class difference, ii) binary cross-entropy loss for AFW layer to emphasize most informative face regions, iii) binary cross-entropy loss for GRU that combined logits, weights and features of each frame of a video to make prediction. Apart from these, author also employed different mechanisms to prevent the network from being overconfident of its predictions.

Similar to this, one of the approaches [121] utilized VGG-16 based Capsule networks with LSTM for spatial and temporal analysis of video sequences. Although the model was

[18] https://github.com/EricGzq/Hybrid-Fake-Face-Dataset

[19] Gated Recurrent Unit

not found much efficient for DFDC dataset as compared to XceptionNet, it has less computational complexity. Later on, a two-branch recurrent network [119] was proposed for deepfake detection. One branch analyzed color domain features while the other focused on amplifying manipulation artifacts using LoG. Combined features from both branches were used to train Bi-LSTM to detect whether a video sequence is fake or real.

Meanwhile, authors of [173] stated that there may be discrepancies among face regions of consecutive frames due to variation in brightness/contrast and size of facial parts. Owing to the fact, they proposed a conv-lstm based residual network (CLRNet). Though the technique showed good generalization ability but performed transfer learning for the same. Since lip-shape tends to exhibit unnatural fluctuations due to deepfaking, a spatio-temporal based network, LipForensics [67], was proposed to detect deepfake. The proposed network was pre-trained for the task of lip-reading first. Only temporal network (Multi Scale Temporal Convolutional Network) was fine-tuned for deepfake detection. Moreover, a simple 3D-CNN based model was utilized in [133] to analyze spatio-temporal features for deepfaked video detection.

Localizing Deepfake Detection Along with the detection of deepfaking in multimedia, some researchers also focused on localizing deepfaked regions. One such approach leveraged a Y-shaped auto-encoder [129] trained in semi-supervised manner. The primary goal of Y-shaped decoder was to perform multi-task learning by segmenting tampered parts using one branch and outputting probability of image spoofing using another. Both branches were made to share their information after each cycle to improve the overall performance. Fine-tuning the auto-encoder with small amount of data further improved performance. Since auto-encoder has ability to cop with unseen attacks, the proposed method was claimed to be generalizable.

Later on, pursuing a similar goal, a group of researchers developed an approach named FakeLocator [75] which analyzed fake texture produced due to upsampling step in GAN. Here, semantic segmentation encoder-decoder architecture was utilized in contrast to [164], wherein small attention maps and incomplete segmentation were not able to point out a fake region. In simple words, a gray-level fakeness prediction map was generated for localizing fake regions by incorporating an attention layer between encoder-decoder architecture. For better generalization, network was trained with diverse variety of deepfakes generated using different GANs. Some researchers utilized dissonance between speech and visual modalities as a detection artifact and proposed a multimodal architecture [36]. This bi-stream network employed 3D-ResNet architecture for visual stream whilst 6-layer CNN model for audio stream. To enforce audio-visual consistency, authors leveraged contrastive loss that effectuated audio-specific and visual specific streams to tie-up. Final loss is combination of binary cross-entropy losses from both streams and contrastive loss, as described by (2). This equation iterates over a number of videos,n where p_i is the prediction for video i and a_i is the actual output. However, p_i^v and p_i^a represents labels predicted by video and audio stream respectively. The author also performed temporal localization to determine which frames are tampered, using Grad-CAM [156].

$$L = \frac{1}{n}[\sum_{i=1}^{n}[(a_i)(d_i^2) + (1-a_i)max(m-d_i,0)^2] - \sum_{i=1}^{n}[(a_i)\log p_i^v + (1-a_i)\log(1-p_i^v)] - \sum_{i=1}^{n}[(a_i)\log p_i^a + (1-a_i)\log(1-p_i^a)]] \tag{2}$$

Content Discrepancies Most of the time, deepfakes has been detected by exploiting dissimilarity or irregularity among content be it audio or video, shape or appearance, behaviour or appearance, fluctuated lip movement. One such approach, proposed in [157], analyzed pixel-wise and region-wise similarity. To perform this, pretrained network, HRNet [166], extracted features in different resolutions. After analyzing pixel-wise similarity through extracted features, all features are fused through spatial attention mechanism to classify manipulated region. Another approach, followed in [124], analyzed the similarity between audio and visual cues with respect to human emotion. Siamese based network architecture [19] with modified triplet loss trained a pair of real video along with its deepfake and provided feature vectors for emotion and modality embedding. As deepfake methods only effect the internal part of face while rest (head, neck, hair) remains unaffected, some researchers analyzed discrepancies among internal face region and its context for deepfake detection [135]. After segmenting facial part and its context using U-Net, two face recognition networks were trained to compute identity embeddings. Comparison of two identity vectors from face and context would determine the presence of tampering.

Meanwhile, deepfake researchers exposed the violation in texture regularity of different facial regions which usually result in high frequency signals. Thereby, they proposed an Invariant Texture Learning Framework (InTeLe) [168] which comprises of an encoder (EfficientNet-b5) and a pair of decoders (U-Net [112]) to study this texture invariance. Since deepfaking cause blending of patches from multiple sources causing inconsistency among different patches of face, patch-wise consistency learning is utilized in [206]. The proposed model utilized ResNet-34 as backend and compared all possible local patches. Recently, another approach [62] analyzed the inconsistency between facial appearance and facial shape of deepfaked faces. The technique utilized 3DMM (3D morphable model) for capturing facial shape information which was compared with facial shape template registered through appearance features. Comparison was done using Mahalanobis distance.

One Class Learning It is pertinent that deep network variants require large amount of training data to make it able to distinguish pristine videos from tampered ones. Collecting large amount of manipulated sequences generated by diverse variety of deepfake generators is much cumbersome, and may become obsolete with the arrival of a new synthesis tool. Owing to the problem, a new perspective was followed in [82] by utilizing only real images(one class) to train a deepfake detector which would treat deepfaked images as anomalies. For this, researchers developed two varaints of variational auto-encoders (VAE). One model is similar to the original VAE [87] with one encoder followed by a decoder that differentiated fake and real videos by computing RMSE between original input to the encoder and reconstructed output from the decoder. By inserting an additional encoder at the end of first model, another model was proposed which computed RMSE between first encoder's input and second encoder's output. Here, second encoder was supposed to extract more efficient features from a reconstructed image if the image is real but the same is not possible from non-real faces. To train this network, a novel loss function was proposed as described in (3). Here, first term demonstrates KL-Divergence while second term is the RMSE between input and output.

$$L = D_{KL}[N(\mu(x), \sigma(x)), N(0, I)] + |X - p(d)| \tag{3}$$

Apart from these deepfake detection techniques, there is a need to focus on multi-face videos/frames which may contain real and fake faces simultaneously. The first paper that addressed this problem utilized a multiple-instance learning (MIL) [108], wherein faces and video were treated as instances and bag respectively. This MIL model was trained to predict a bag label based on spatial and temporal embeddings of different instances present in the bag. Pretrained XcpetionNet was utilized for spatial feature extraction. Since new deepfake generation methods are continuously emerging, it is important to develop a network that can cope with future deepfakes also. Owing to this issue, a new forensic framework [101] was developed by deepfake researchers based on encoder-decoder architecture. After training the proposed model on one type of deepfake, the idea is to fine-tune the same on few frames of new generation deepfakes. Though fine-tuning seems impractical for real-life scenarios, it can work well even if small amount of data is available. A completely different approach [84] utilized an ensemble of VGG16, IncpetionV3 and XceptionNet for the task of deepfake detection. Ensemble model helps to protect against adversarial attack.

3.2.3 Biological signals for deepfake detection

Almost every technique proposed in literature depends on facial features of a person present in video; this constraint made every technique dependant on video content. Contrary to this, another technique [39] exploited heart rate of a person which was not supposed to be replicated efficiently in the fake video. This heart rate was extracted using PPG (PhotoPlethysmoGraphy) signal, through six facial features, that is more robust in dynamic scenarios as compared to other signals. Later on, it was analyzed that prediction of heart rate using PPG from video is somewhat complex and time-consuming. In view of this, Neural ODE (Neural Ordinary Differential Equations) [28], which is supposed to model time series data of heart rate, was utilized for heart rate prediction to detect deepfaked videos [56]. Here, Neural ODE was trained using blood flow caused skin-color variation, forehead's optical intensity, and temporal changes in color.

Motivated by biological signals, some other researchers leveraged convolutional network with attention mechanism (CAN) [72], already pretrained based on rPPG (remote PPG) features for heart rate estimation. CAN is composed of two parallel CNN's which tends to exploit spatial and temporal information from suspected video. Meanwhile, another approach focused on analyzing heartbeat rhythms using a similar spatio-temporal-attention concept. Rather than raw faces, motion magnified spatio-temporal (MMST) representations were used to train spatial and temporal networks which outputs spatial and temporal attention respectively. Final classification was done by training ResNet18 with attentional MMST map.

3.2.4 Blockchaining smart contract

Haya and khaled provided a new way [68] to combat deepfake videos that, rather than analyzing video content, created a smart contract, called Ethereum smart contract for every video on its creation. Smart contract provides an ability to each video to be traced back to its original source. The proposed system was named as decentralized PoA (Proof of Authenticity) system which utilized blockchaining to provide details of transactions performed on data. According to this technique, metadata associated with a video contains the address of smart contract and its creator which would be stored in IPFS (InterPlanetary File System). IPFS, in return, generates a unique hash value for each video that serves as an address of files and metadata of the respective video. Any post-production manipulation on video will require the source's permission and if permission granted, create a child contract linked to a parent contract using blockchain. This technique has not been implemented yet but may be able to work well on deepfake videos. If automated, it will be able to ensure integrity, non-repudiation, and authorization of any type of digital content.

To conclude, it is important to note that technology can do both good and evil, so as deepfakes. In spite of using deepfake technology to create inappropriate content and infringe privacy rights, it can be used as a face de-identification method to protect privacy in various medical and other private issues. In contrast to previous face de-identification methods like masking, blurring, or pixelize face, deepfakes have been proved to retain original face information [208]. Despite this, the risk of misusing the identity of any person by deepfaking it on another person is increasing day-by-day with improvements in previous techniques. Various deep learning techniques have been developed for spotting deepfaked images and videos but the generalizing ability of most of the techniques is questionable. If one depends on the training dataset utilized [169], the other technique developed to counter this dependency becomes dependant on blending artifacts [105]. Though some approaches were tried to improve the performance on unknown data [57, 75, 82, 85, 189] but requires more focus. Moreover, most of the techniques were evaluated on early generation deepfakes, which exhibit easily-recognizable visual inconsistencies. These techniques should also be evaluated on more advanced deepfaked datasets, DeeperForensics-1.0 [77] and DFDC containing diverse variety of videos with much realistic fake content. In view of the state, some researchers also tried to defend against deepfake attacks [68]. But, it requires specialized devices and software that can't be made available to everyone. In the wake of improving deepfake attacks, such defense procedures would not be fruitful for long-term.

3.3 Lip-sync deepfake detection

Unlike deepfaking the whole face, lip-sync deepfakes require only mouth and lip area to be manipulated as discussed in Section 1.2. Analysis of facial features or face alignment can't possibly detect lip-sync deepfake where only lip movement was adjusted with respect to a specific audio clip. The idea of lip-sync deepfake detection was raised in [5] where along with other facial features for face swap detection, researchers also analyzed lip raiser and lip corner features to compute the distance between different corners of lips. This technique trained a machine learning classifier, SVM, using facial and head movement features extracted using the OpenFace2 toolkit. However, eye blink feature out of 17 features provided by the toolkit was eliminated due to its non-uniqueness. Along with these, mouth stretch and lip suck features were also extracted, creating a set of 20 features in total. 190 combinations of these features created a feature vector of dimension 190. These 190-D features of POI's (Person Of Interest) of different persons, when visualized in 2-D, were well separated from each other which proved the uniqueness of the head and face movements of different persons. However, the network needs to be trained only with authentic videos of a particular person to learn only original features so that fake features can be differentiated.

Though not using deep neural networks, techniques of meddling different audio track and synchronizing someone's lips accordingly in a video have been evolving for many years. Thereby, many methods were proposed in literature to detect inconsistent synchronization between lip movement and audio track present in a video sequence. Lip de-synchronization caused due to deep neural networks also can be evaluated using these methods. For this, one of the authors proposed a two-stream ConvNet architecture which was trained end-to-end using speech and mouth images mapping [38]. Mouth region image from the face and MFCC values from audio data were fed into two separate models; VGG-M [37] and LSTM model [24]. Unlike previous methods, models were trained using un-annotated and even noisy data also.

Following a similar idea, facial landmarks specifically from the lip area were detected from the video part and MFCC (Mel-Frequency Cepstral Coefficients) features from the

Table 15 Lip-sync Deepfake detection techniques (A: Accuracy, EER: Effective Error Rate)

Technique	Dataset	Results	Performance
SVM trained using head-movement features [5]	Youtube videos Forged using face swap, lip-sync and puppet master	Average A: 95%	Robust against compression and video length. Trained using original videos only. Effective only if person is facing the camera. Ineffective if video portrays less known person.
Two-stream ConvNet architecture [38]	Columbia [21] OuluVS2 [9]	A:>99%	No annotated data required. Tolerant to noise
Mouth's landmark and audio's MFCC feature vector [95] GMM, SVM, MLP and LSTM Classifier	Created from: VidTIMIT [152], AMI [15] and GRID [143]	Average EER: 24.24% (LSTM)	Generated a new deepfake dataset Only front-face videos considered. Large database required.
Mouth's landmark (Visual) MFCC and DNN embeddings (Audio) [93] LSTM classifier	VidTIMIT [152] AMI [15] GRID [143]	Average EER: 24% (Known data) Average EER: 26.9% (Unknown data)	Higher error rate on unseen data.
Aural and oral dynamics [4]	Generated and collected	A: 90%	Need reference data. Ear must be visible.

audio part [95]. Combined features from both were fed to GMM (Gaussian Mixture Model), SVM (Support Vector Machine), MLP (Multilayer Perceptron), and LSTM (Long Short-Term Memory) for training. However, best results were found by applying LSTM classifier especially on GRID dataset due to a large amount of training data available with an Effective Error Rate (EER) of 14.12% during testing. Extending this technique, some researchers [93] utilized a different set of audio features for deepfaked audio detection. To extract audio feature set, along with MFCC, embeddings of Deep Neural Network trained for speech recognition were utilized. After training different LSTM classifiers using audio and visual features, the technique provided a lowest EER of 4.5% on GRID dataset. However, the technique was not found effective for AMI dataset as it contains diverse variety of videos.

Deepfaking, be it face swap or lip-sync, synthesize only facial area and decouple rest of the part such as ear which can be provide a significant biometric identification of certain individual, Deepfake researchers analyzed the dynamics of ear motion which are supposed to be different from mouth and jaw motion [4]. To determine whether a video has been deepfaked, correlation between aural (ear motion) and oral (audio RMSE[20] and distance between lips) dynamics is measured. LR model was trained using twelve such correlations. As aural biometric, authors compared shape of ear from suspected video with reference shape.

Apart from these detection techniques, described in Table 15, lip-sync deepfake and puppetry deepfake was not focussed much because there are no public datasets available for

[20] Root Mean Square Energy

both, and deepfakes generated using these are also not much convincing. But increasing advancement in the deep neural networks and computer graphic techniques has escalated the risk of generating convincing full body deepfakes. After video deepfake, the next risk of misinformation has started with the emergence of deepfaking in audio through which a person can deepfake one's voice with other person's. Even thought of such a great technology inculcates fear in society; any person's voice can be misused now. A person's voice is a part of his identity, but applications like Lyrebird[21] are easily faking one's identity using his/her voice. The technology has also incorporated into the Podcasting software 'Descript' which after learning a person's voice from a small dataset can ape that same voice.

3.4 Audio deepfake detection

Since advent recently, techniques for detecting audio deepfakes have not been focused much. Considering the importance of defending society from fraudsters, a challenge on ASVspoof[22] was launched in 2019 for spoofing attack detection. To evaluate submitted techniques, organizers also released a dataset containing audio sequences synthesized using 17 TTS and VC techniques. Besides, a tool named 'Resemblyzer'[23] has been made public for the detection of AI-synthesized fake audio, which utilized a voice encoder [185] that was originally designed for speaker identification.

To combat deepfake audio, some researchers analyzed higher-order correlation of audio using bispectral analysis [153]. Depending on the fact that simple decomposable signals often produce glaring artifacts in bi-coherence, it was considered as a fingerprint to distinguish real audio from AI-synthesized one. To counter the effect of noise present in bi-coherence, it was averaged across multiple waveforms. Training SVM classifier using these correlation features provided an average accuracy of 95% on audio sequences synthesized using DC-TTS [170] and Tacotron2 [158]. However, evaluation of these fake audios through human auditory system revealed that humans can't differentiate it accurately due to continuous exposure to natural speech and lack of micro-level analysis using experimental instrumentation. Another work performed by researchers, in the direction of audio deepfake detection, was by using spectrograms of the audio clips [46], which provide visual representation of these audio clips. These spectrograms were supposed to be different for original and fake video and hence, used to train the detector model which performed convolutions of mel-spectrograms over time dimension. After evaluating the model on Google's 2019 AVSSpoof dataset, 90% of the test audios were correctly predicted. These techniques have initiated a path to deepfake audio detection and can be improvised further to detect future audio deepfakes that are on the verge of being developed.

To limit the dependency on synthesis traces, as in [153], another deep-learning model, DeepSonar [188], analyzed the pattern of neuron activation of each layer. Real and fake voices are supposed to have different neuron coverage patterns. A pre-trained thin-ResNet [199] SR (Speech Recognition) model was utilized as a backend. Two neuron coverage criteria were utilized where one counted number of activated neurons (having a value greater than threshold) in each layer while the other extracted top 5 neurons from each layer. The former one was termed ACN (Average Count Neuron) while the later one TKAN (Top-k Activated Neuron). Feature vectors computed using these criteria were used to train a

[21] Available at: https://www.descript.com/lyrebird-ai?source=lyrebird

[22] Available at: https://www.asvspoof.org/

[23] Available at: https://github.com/resemble-ai/Resemblyzer

Table 16 Audio deepfake detection techniques (A: Accuracy, EER: Effective Error Rate)

Technique	Dataset	Results	Performance
SVM trained using correlation features of audio waveforms [153]	1800 samples from LJ speech dataset (synthesized using DC-TTS [170] and Tacotron2 [158]	Average A: 95%	Handcrafted features becomes obsolete for advanced deepfakes.
DeepSonar (Analyzed neuron activation behaviour) [188]	Publically available: TTS synthesized [10] Self created using: Sprocket [90] Baidu TTS [13]	A: 98.1% EER: 2%	Performance detrirorates on resampling and noise attacks. adversarial attacks might change neuron activation behaviour.
ResNet with FreqAugment layer and LMCL function [27]	ASVSpoof 2019 [178]	EER: 1.26%	Good performance. More focused on spoofing attack detection.
Conv+Bi-LSTM [31]	ASVSpoof 2019 [178] FF++ [148] Celeb-DF [110]	EER: 4.17% A: 100% A: 100%	Good performance.

shallow neural network of five fully connected layers. The proposed model was tested on diverse varieties of audio sequences recorded using different languages, synthesized either by VC and TTS approach (Table 16).

Despite the development of various techniques for the ASVspoof 2019 competition [178], there is always a need for more generalized detection techniques. Focusing on a similar motive, researchers in [27] proposed a DNN framework using LMCL (Large Margin Cosine Loss) function that minimizes intra-class variance and maximizes inter-class variance. LFB (Linear Filter Bank) features, computed from raw audio, were first fed into the FreqAugment layer to randomly drop-out adjacent frequency channels and then trained a Residual Network (ResNet). For robust performance, training data was imposed with a variety of augmentations and were re-recorded through VoIP channel to simulate spoofing attacks. Focusing on same goal, one of the researcher proposed temporal architectures for visual deepfake detection and for audio spoof detection. XceptionNet was combined with Bi-LSTM for visual deepfake while for audio spoof detection, a combination of 4 convolutional layers and Bi-LSTM layer was leveraged. However, the method performed better with Kullback-Leibler (KL) divergence loss function.

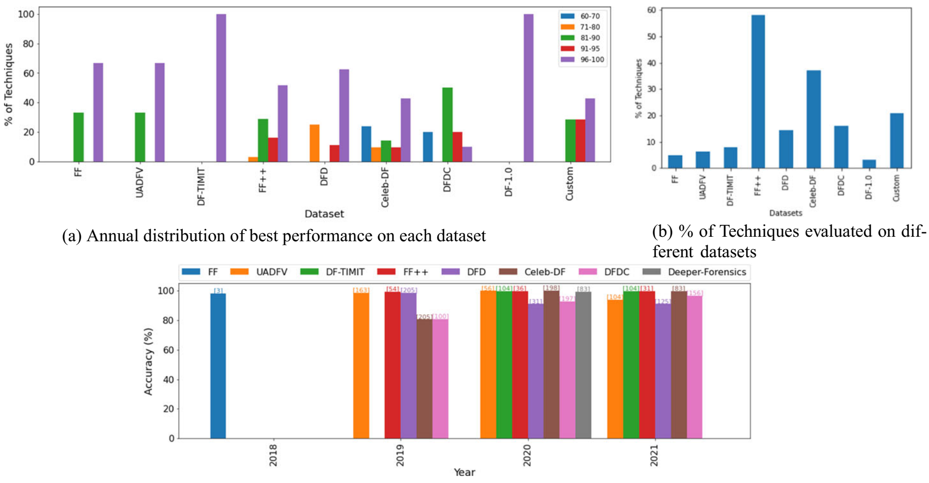
4 Discussion and future scope

Research in multimedia forensics has been greatly progressed since its inception. As discussed in this survey, last three years witnessed a substantial improvement in deepfake detection approaches. Table 17 provides summary of different categories of deepfake detection approaches from state-of-art. On the inception of deepfakes, forensic researchers started proposing approaches based on handcrafted features and deep networks. As shown in Table 17, 75% of techniques proposed in first and second year of deepfake detection research utilized handcrafted features. On the contrary, with advancement in AI technologies in 2020, 80% of deepfake detection research was focused on different deep network

Table 17 Summary of Deepfake Detection Techniques (Maximum performance of all techniques is provided)

		Technique			Analysis			Tampering						Evaluation Dataset												Performance		
Year	Paper	Feature +SN	Feature + DNN	Deep Network	Spatial	Spatio-Temporal	Audio	Face replacement	Face Attribute	Face re-enactment	Lip-sync	Audio	Model	UADFV	DF-TIMIT	FF++	DFD	Celeb-DF	DFDC	Deeper-Forensics	Others	Custom	Generalization	Explainability	Localization	Accuracy	AUC	EER
2018	[109]		●			●		●		●			VGG16+LSTM									●					99	
	[65]		●			●		●		●			Inceptionv3+LSTM									●				97.2		
	[2]			●				●		●			Inception			●						●				98.4		
	[93]		●			●	●				●		SVM								●							14.12
2019	[150]		●			●		●		●			DenseNet			●										96.9		
	[201]	●			●			●					SVM	●												89		
	[120]	●								●			LR			●											86.6	
	[95]		●			●	●				●		LSTM								●	●						4.5
	[8]		●			●				●			VGG			●										81.61		
	[53]	●						●					SVM				●									87		
	[105]		●		●			●		●			HRNet			●	●	●	●				●	●			98.52	
	[130]			●	●			●		●			VGG16			●										99.37		
	[164]			●	●			●		●			CNN (Attention)	●				●				●		●			98.4	
	[57]			●	●			●		●			CNN+HMN			●						●	●			99.43		
	[129]			●	●			●		●			AutoEncoder			●							●		●	92.50		8.07
	[39]		●		●			●		●			CNN			●							●			91.07		
	[56]		●			●		●					Neural ODE		●													1.54
	[5]	●			●			●		●	●		SVM									●					99	
	[95]		●			●	●				●		LSTM								●							4.5
	[153]	●				●	●					●	SVM								●					95		
2020	[99]		●		●					●			ResNet			●										99.96		
	[63]	●							●				KNN									●				90.22		
	[52]	●						●					SVM				●									90		
	[76]			●	●			●		●			MobileNetv3			●						●				97.02		
	[189]			●	●			●	●				ResNet50			●		●	●			●				98.5		
	[96]			●	●			●		●			Facenet			●		●									99.2	
	[17]			●	●			●		●			EfficientNetB4			●			●								94.440.3294	
	[3]			●		●		●		●			VGG16			●	●	●									98.9	
	[126]			●		●		●					EfficientNetb5+GRU						●							92.61		0.321
	[121]			●		●		●					VGG16+LSTM						●							83.42		
	[66]			●	●			●	●	●			AREN			●						●				98.52		
	[124]			●	●		●	●					OpenFace+MFCC		●				●							96.3		
	[75]			●					●				AutoEncoder									●				99.85		
	[36]			●	●		●	●					CNN+LSTM		●				●						●		97.9	
	[82]			●	●			●		●			VAE			●	●									98.20		
	[119]			●		●		●		●			Bi-LSTM			●		●								96.43		
	[108]			●		●		●		●			XceptionNet			●		●								99.23		
	[188]			●			●					●	ResNet								●					93.1		
	[27]		●	●	●							●	ResNet								●							
	[135]			●	●			●			●		Xception variant			●		●					●			75		
	[173]			●		●		●		●			Conv+LSTM			●	●						●			99.35		
	[168]			●	●			●					Encoder-Decoder			●		●									95.5	
	[55]			●	●			●		●			Modified Xception	●		●		●					●				99.9	
	[42]			●		●		●		●			Temporal-ID + 3DMM			●	●	●								90		
	[206]			●	●					●	●		ResNet-34			●	●	●	●	●							99.79	
	[67]			●		●		●		●			Resnet-3D +MSTCN			●		●	●	●			●			97.6		
	[45]		●			●		●		●			3-Dense-layers			●	●	●								89.79		
	[131]		●			●		●					ResNet					●									87.9	
	[31]			●		●	●	●		●		●	Conv+LSTM			●		●			●		●			100		4.17
	[72]			●		●		●					Attention					●								98.7		
	[146]			●		●		●		●			Attention			●							●			98		
2021	[20]		●		●			●		●			ResNet50			●							●			97		
	[92]		●		●			●		●			CNN			●		●						●		85.24		
	[133]			●		●		●		●			3D-CNN		●			●								99.33		
	[83]		●		●			●					HRNet					●					●			86		
	[101]			●	●			●		●			AutoEncoder			●					●		●			99.8		
	[62]			●	●			●					3DMM			●										87.3		
	[157]			●	●			●		●			HRNet			●		●						●	●	99.80		
	[84]			●	●			●					VGG+Inception+Xception		●				●							99.68		
	[198]			●	●			●		●			Transformer	●		●	●									93.75%		
	[4]		●			●		●		●	●		Logistic Regression								●					98%		
	[186]			●	●			●		●			Transformer			●		●				●	●		●	99.5%		

variants. The reason for this deviation is that handcrafted features become obsolete with time due to advancement in generation techniques. Figure 9c also suggests that large proportion of best performing techniques falls under category of deep network variants. However, advancement in deepfake generation techniques indulge more challenges for deepfake detectors and thereby, new approaches for deepfake detection would always be a necessity. From the state-of-art survey performed in this paper, there are various issues that need to addressed for future research.



(a) Annual distribution of best performance on each dataset

(b) % of Techniques evaluated on different datasets

(c) Accuracy range of techniques on specific dataset

Fig. 9 Performance visualization of techniques on different datasets

- **Lack of experimentation:** First issue that needs to be taken care of is lack of experimentation on more realistic deepfakes. Thorough analysis of state-of-art papers revealed that only 3% of existing approaches evaluated their performance on Deeper-Forensics dataset and 16% on DFDC, both of which exhibit diverse content and deepfakes generated through advanced generators. On the other hand, 58% of researchers reported their evaluation on FF++ dataset, deepfake videos of which contain visual tampering artefacts. Figure 9b shows the proportion of techniques evaluated on each dataset. As demonstrated in Fig. 9a, most of the techniques evaluated on DFDC dataset provided accuracy between 81-90% while on other datasets, maximum accuracy obtained was greater than 95%. All these factors inculcate the need of more experimentations on current generation datasets and proposing more efficient solutions for deepfake detection.
- **Generalizability:** Another major issue is generalizability on different datasets. Analysis provided in Fig. 9c demonstrates that remarkable performance has been obtained on every dataset through one or several techniques. Despite such interesting performance, it must be considered that most of such techniques didn't even evaluate their performance on other datasets. However, some researchers have tried to generalize their technique on different available datasets [57, 75, 85, 189] but were found dependant in one way or other. Such approaches were either content-specific or generator-specific which are likely to fail soon due to the continuous evolvement of deepfake generating techniques. These dependant models tend to overfit the training data and thereby might not generalize on new data. Therefore, there is a need of new proposals that can ensure good generalization which calls for extensive validation of detectors on diverse range of datasets and different types of manipulations. Moreover, a technique must be designed that could be able to face the real challenge of deepfake detection independent of generation techniques utilized.
- **Explainability and tampering localization:** Future research should also focus on explainability of deepfake detection approaches.Although various deep learning models [2, 65,

109, 130] have proven best for detection of deepfake video, these follow black box architecture and are unexplainable. An explainable model may allow us to improve the design of network and provides higher robustness against diverse range of manipulations. However, generation of whole-body deepfakes and lip-sync deepfakes calls for better localization of tampering by deepfake detectors which has not been focused much.

- **Lack of lip-sync and audio deepfake datasets:** Although lip-sync and audio deepfakes are more threatening than face swap deepfakes, no sufficient detection procedures are available for these in literature. One of the reason for this is the lack of datasets for the same. Moreover, with continuing advancement in deepfake generation procedures, existing datasets and approaches evaluated on those become obsolete. Thereby, in future, more diverse datasets should be designed to evaluate the robustness of different techniques.
- **Emerging possibilities in new domain:** At last, it is recommended to use active forgery detection schemes by pre-embedding information like watermarks or smart contracts on the content at recording time. Although complicated, research can also be directed towards the idea of tracing all the processing operations performed on the video using blockchain [68]. Deepfake forgery detection may also benefit from recent trends and advances in deep learning procedures such as self-supervised learning and attention modeling.

5 Conclusion

In the era of digital world capable of providing reliable evidence of any situation, a trouble-free creation of tampered videos made it absurd to trust its content. Advancement in AI has given impulse to media manipulators, increasing the online availability of deepfakes by more than 100% from last year [6]. Usage of AI-based deepfaking applications for revenge porn, bullying, fabricating video evidence or news, blackmailing, or political sabotage, can make the life of targeted person hard. With the advent of applications like ‘Deepnude’ which can create and spread fake nudes, the dignity of women is at stake. Audio deepfakes has also made people able to talk anonymously to anyone. Due to the two-player nature of this research field, the advent of AI has also contributed in the detection of manipulated content. But the quality of deepfake videos is continuously improving with AI advancement; new challenges appear everyday. To cope with unforeseen menaces, a more efficient detection method will always be a necessity.

This survey paper presents an analysis of visual media tampering detection techniques with special emphasis on deepfake detection. However, the analysis is not restricted to facial manipulation detection but approaches developed for lip-sync deepfake detection and audio deepfake detection were also taken into consideration. At the end, a critical summarization of state-of-art approaches is provided with an analysis of best performing techniques on different datasets. Nevertheless, techniques for detecting tampered media and more specifically deepfakes only provides a partial solution to the problem at hand. Along with improvisation in detection methods, some authentic laws should be formulated that can restrict the development and usage of such user-friendly but malicious tools creating deepfakes. Of course, digital content has been faked before, either in a useful or malicious manner, but the level of realism in fake content has increased due to deep learning technologies. There are no signs of deepfake technology being slowed down, so researchers and government should come forward jointly to discuss the issue and fight against this devastating technology.

Acknowledgments This Work is carried out at Design Innovation Center, Panjab University, Chandigarh, INDIA, established by the Ministry of Education, Government of India.

References

1. Adami N, Signoroni A, Leonardi R (2007) State-of-the-art and trends in scalable video compression with wavelet-based approaches. IEEE Trans Circ Syst Video Technol 17(9):1238–1255
2. 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). IEEE, pp 1–7
3. Agarwal S, El-Gaaly T, Farid H, Lim SN (2020) Detecting deep-fake videos from appearance and behavior. arXiv:2004.14491
4. Agarwal S, Farid H (2021) Detecting deep-fake videos from aural and oral dynamics
5. Agarwal S, Farid H, Gu Y, He M, Nagano K, Li H (2019) Protecting world leaders against deep fakes. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 38–45
6. Ajder H Deepfake threat intelligence: a statistics snapshot from june 2020. http://deeptracelabs.com/deepfake-threat-intelligence-a-statistics-snapshot-from-june-2020/
7. Al-Sanjary OI, Ahmed AA, Sulong G (2016) Development of a video tampering dataset for forensic investigation. Forensic Sci Int 266:565–572
8. Amerini I, Galteri L, Caldelli R, Del Bimbo A (2019) Deepfake video detection through optical flow based cnn. In: Proceedings of the IEEE international conference on computer vision workshops, pp 0–0
9. Anina I, Zhou Z, Zhao G, Pietikäinen M (2015) Ouluvs2: a multi-view audiovisual database for non-rigid mouth motion analysis. In: 2015 11Th IEEE international conference and workshops on automatic face and gesture recognition (FG), vol 1. IEEE, pp 1–5
10. APTLY: Audio processing techniques lab at york. http://bil.eecs.yorku.ca/aptly-lab./
11. Aslani S, Mahdavi-Nasab H (2013) Optical flow based moving object detection and tracking for traffic surveillance. Int J Electr Comput Eng 7(9):1252–1256
12. Baddar WJ, Gu G, Lee S, Ro YM (2017) Dynamics transfer gan:, Generating video by transferring arbitrary temporal dynamics from a source video to a single target image. Accessed 5 May 2021. arXiv:1712.03534
13. Baidu text-to-speech system. https://cloud.baidu.com/product/speech/tts
14. Baltrušaitis T, Robinson P, Morency LP (2016) Openface: an open source facial behavior analysis toolkit. In: 2016 IEEE Winter conference on applications of computer vision (WACV). IEEE, pp 1–10
15. Barker J (2013) The grid audiovisual sentence corpus, available at: http://spandh.dcs.shef.ac.uk/gridcorpus/
16. Bidokhti A, Ghaemmaghami S (2015) Detection of regional copy/move forgery in mpeg videos using optical flow. In: 2015 The international symposium on artificial intelligence and signal processing (AISP). IEEE, pp 13–17
17. Bonettini N, Cannas ED, Mandelli S, Bondi L, Bestagini P, Tubaro S (2020)
18. Bregler C, Covell M, Slaney M (1997) Video rewrite: Driving visual speech with audio. In: Proceedings of the 24th annual conference on Computer graphics and interactive techniques, pp 353–360
19. Bromley J, Guyon I, LeCun Y, Säckinger E, Shah R (1994) Signature verification using a” siamese” time delay neural network. In: Advances in neural information processing systems, pp 737–744
20. Caldelli R, Galteri L, Amerini I, Del Bimbo A (2021) Optical flow based cnn for detection of unlearnt deepfake manipulations. Pattern Recogn Lett 146:31–37
21. Chakravarty P, Tuytelaars T (2016) Cross-modal supervision for learning active speaker detection in video. In: European conference on computer vision. Springer, pp 285–301
22. Chan C, Ginosar S, Zhou T, Efros AA (2019) Everybody dance now. In: Proceedings of the IEEE international conference on computer vision, pp 5933–5942
23. Chao J, Jiang X, Sun T (2012) A novel video inter-frame forgery model detection scheme based on optical flow consistency. In: International workshop on digital watermarking. Springer, pp 267–281
24. Chatfield K, Simonyan K, Vedaldi A, Zisserman A (2014) Return of the devil in the details:, Delving deep into convolutional nets. arXiv:1405.3531
25. Chen H, Chandrasekar V, Tan H, Cifelli R (2019) Rainfall estimation from ground radar and trmm precipitation radar using hybrid deep neural networks. Geophysical Research Letters

26. Chen H, Wo Y, Han G (2018) Multi-granularity geometrically robust video hashing for tampering detection. Multimed Tools Appl 77(5):5303–5321
27. Chen T, Kumar A, Nagarsheth P, Sivaraman G, Khoury E (2020) Generalization of audio deepfake detection. In: Proceedings of the Odyssey 2020 the speaker and language recognition workshop, pp 132–137
28. Chen TQ, Rubanova Y, Bettencourt J, Duvenaud D. K (2018) Neural ordinary differential equations. In: Advances in neural information processing systems, pp 6571–6583
29. Cheung GK, Baker S, Hodgins J, Kanade T (2004) Markerless human motion transfer. In: Proceedings of the 2nd international symposium on 3d data processing, visualization and transmission, 2004. 3DPVT 2004. IEEE, pp 373–378
30. Chingovska I, Anjos A, Marcel S (2012) On the effectiveness of local binary patterns in face antispoofing. In: 2012 BIOSIG-proceedings of the international conference of biometrics special interest group (BIOSIG). IEEE, pp 1–7
31. Chintha A, Thai B, Sohrawardi SJ, Bhatt K, Hickerson A, Wright M, Ptucha R (2020) Recurrent convolutional structures for audio spoof and video deepfake detection. IEEE J Sel Top Signal Process 14(5):1024–1037
32. Cho W, Choi S, Park D. K, Shin I, Choo J (2019) Image-to-image translation via group-wise deep whitening-and-coloring transformation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 10639–10647
33. Choi Y, Choi M, Kim M, Ha JW, Kim S, Choo J (2018) Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8789–8797
34. Choi Y, Choi M, Kim M, Ha JW, Kim S, Choo J (2018) Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8789–8797
35. Chollet F (2017) Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1251–1258
36. Chugh K, Gupta P, Dhall A, Subramanian R (2020)
37. Chung JS, Zisserman A (2016) Lip reading in the wild. In: Asian conference on computer vision. Springer, pp 87–103
38. Chung JS, Zisserman A (2016) Out of time: automated lip sync in the wild. In: Asian conference on computer vision. Springer, pp 251–263
39. Ciftci UA, Demir I (2019) Fakecatcher:, Detection of synthetic portrait videos using biological signals. arXiv:1901.02212
40. Cole S (2017) Ai-assisted fake porn is here and we're all fucked https://www.vice.com/en_us/article/gydydm/gal-gadot-fake-ai-porn
41. collection, D.: Xiph.org video test media. Accessed 5 May 2021. https://media.xiph.org/video/derf/
42. Cozzolino D, Rössler A, Thies J, Nießner M, Verdoliva L (2020) Id-reveal:, Identity-aware deepfake video detection. arXiv:2012.02512
43. D'Amiano L, Cozzolino D, Poggi G, Verdoliva L (2018) A patchmatch-based dense-field algorithm for video copy–move detection and localization. IEEE Trans Circ Syst Video Technol 29(3):669–682
44. De Roover C, De Vleeschouwer C, Lefebvre F, Macq B (2005) Robust video hashing based on radial projections of key frames. IEEE Trans Signal Process 53(10):4020–4037
45. Demir I, Ciftci UA (2021) Where do deep fakes look? synthetic face detection via gaze tracking. arXiv:2101.01165
46. (2019) Dessa: Detecting audio deepfakes with ai. available at:. https://medium.com/dessa-news/detecting-audio-deepfakes-f2edfd8e2b35
47. Ding X, Zhang D (2019) Detection of motion-compensated frame-rate up-conversion via optical flow-based prediction residue. Optik p 163766
48. Dolhansky B, Bitton J, Pflaum B, Lu J, Howes R, Wang M, Ferrer CC (2020) The deepfake detection challenge dataset. arXiv:2006.07397
49. Dolhansky B, Howes R, Pflaum B, Baram N, Ferrer CC (2019) The deepfake detection challenge (dfdc) preview dataset. arXiv:1910.08854
50. Dong Q, Yang G, Zhu N (2012) A mcea based passive forensics scheme for detecting frame-based video tampering. Digit Investig 9(2):151–159
51. Dufour N (2019) Google ai blog. contributing data to deepfake detection research. Accessed 5 May 2021. https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html
52. Durall R, Keuper M, Keuper J (2020) Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 7890–7899

53. Durall R, Keuper M, Pfreundt F. J, Keuper J (2019) Unmasking deepfakes with simple features. arXiv:1911.00686
54. Esser P, Haux J, Milbich T et al (2018) Towards learning a realistic rendering of human behavior. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 0–0
55. Feng D, Lu X, Lin X (2020) Deep detection for face manipulation. In: International conference on neural information processing. Springer, pp 316–323
56. Fernandes S, Raj S, Ortiz E, Vintila I, Salter M, Urosevic G, Jha S (2019) Predicting heart rate variations of deepfake videos using neural ode. In: Proceedings of the IEEE international conference on computer vision workshops, pp 0–0
57. Fernando T, Fookes C, Denman S, Sridharan S (2019) Exploiting human social cognition for the detection of fake and fraudulent faces via memory networks. arXiv:1911.07844
58. Garg R, Varna AL, Hajj-Ahmad A, Wu M (2013) "seeing" enf: power-signature-based timestamp for digital multimedia via optical sensing and signal processing. IEEE Trans Inf Forensics Secur 8(9):1417–1432
59. Garrido P, Valgaerts L, Sarmadi H, Steiner I, Varanasi K, Perez P, Theobalt C (2015) Vdub: Modifying face video of actors for plausible visual alignment to a dubbed audio track. In: Computer graphics forum, vol 34. Wiley Online Library, pp 193–204
60. Grisham S (2018) Stephanie grisham on twitter. tampering performed on white house secretary's video https://twitter.com/PressSec/status/1060374680991883265
61. Guan H, Kozak M, Robertson E, Lee Y, Yates AN, Delgado A, Zhou D, Kheyrkhah T, Smith J, Fiscus J (2019) Mfc datasets: Large-scale benchmark datasets for media forensic challenge evaluation. In: 2019 IEEE Winter applications of computer vision workshops (WACVW). IEEE, pp 63–72
62. Guan W, Wang W, Dong J, Peng B, Tan T (2021) Robust face-swap detection based on 3d facial shape information. arXiv:2104.13665
63. Guarnera L, Giudice O, Battiato S (2020) Deepfake detection by analyzing convolutional traces. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, pp 666–667
64. Güera D, Baireddy S, Bestagini P, Tubaro S, Delp EJ (2019) We need no pixels:, Video manipulation detection using stream descriptors. arXiv:1906.08743
65. 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). IEEE, pp 1–6
66. Guo Z, Yang G, Chen J, Sun X (2020) Fake face detection via adaptive residuals extraction network. arXiv:2005.04945
67. Haliassos A, Vougioukas K, Petridis S, Pantic M (2020) Lips don't lie:, A generalisable and robust approach to face forgery detection. arXiv:2012.07657
68. Hasan HR, Salah K (2019) Combating deepfake videos using blockchain and smart contracts. IEEE Access 7:41596–41606
69. He Z, Zuo W, Kan M, Shan S, Chen X (2019) Attgan: Facial attribute editing by only changing what you want. IEEE Trans Image Process 28(11):5464–5478
70. He Z, Zuo W, Kan M, Shan S, Chen X (2019) Attgan: Facial attribute editing by only changing what you want. IEEE Trans Image Process 28(11):5464–5478
71. Hecker C, Raabe B, Enslow R. W, DeWeese J, Maynard J, van Prooijen K (2008) Real-time motion retargeting to highly varied user-created morphologies. ACM Transactions on Graphics (TOG) 27(3):1–11
72. Hernandez-Ortega J, Tolosana R, Fierrez J, Morales A (2020) Deepfakeson-phys:, Deepfakes detection based on heart rate estimation. arXiv:2010.00400
73. Horn BK, Schunck BG (1981) Determining optical flow. Artificial intelligence 17(1–3):185–203
74. Hsieh CK, Chiu CC, Su PC (2018) Video forensics for detecting shot manipulation using the information of deblocking filtering. In: 2018 IEEE 42Nd annual computer software and applications conference (COMPSAC), vol 2. IEEE, pp 353–358
75. Huang Y, Juefei-Xu F, Wang R, Xie X, Ma L, Li J, Miao W, Liu Y, Pu G (2020) Fakelocator:, Robust localization of gan-based face manipulations via semantic segmentation networks with bells and whistles. arXiv:2001.09598
76. Jeon H, Bang Y, Woo SS (2020) Fdftnet:, Facing off fake images using fake detection fine-tuning network. arXiv:2001.01265
77. Jiang L, Wu W, Li R, Qian C, Loy CC (2020) Deeperforensics-1.0:, A large-scale dataset for real-world face forgery detection. arXiv:2001.03024
78. Jr EO (2019) Thieves used audio deepfake of a ceo to steal $243,000 https://www.vice.com/en_in/article/d3a7qa/thieves-used-audio-deep-fake-of-a-ceo-to-steal-dollar243000

79. Karras T, Aila T, Laine S, Lehtinen J (2017) Progressive growing of gans for improved quality, stability, and variation. arXiv:1710.10196
80. Karras T, Laine S, Aila T (2019) A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4401–4410
81. Karras T, Laine S, Aittala M, Hellsten J, Lehtinen J, Aila T (2019) Analyzing and improving the image quality of stylegan. arXiv:1912.04958
82. Khalid H, Woo SS (2020) Oc-fakedect: Classifying deepfakes using one-class variational autoencoder. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, pp 656–657
83. Khalil SS, Youssef SM, Saleh SN (2021) icaps-dfake: an integrated capsule-based model for deepfake image and video detection. Future Internet 13(4):93
84. Khan SA, Artusi A, Dai H (2021)
85. 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). IEEE, pp 1–6
86. Kingma DP, Dhariwal P (2018) Glow: Generative flow with invertible 1x1 convolutions. In: Advances in neural information processing systems, pp 10215–10224
87. Kingma DP, Welling M (2013) Auto-encoding variational bayes. arXiv:1312.6114
88. Kingra S, Aggarwal N, Singh R. D (2016) Video inter-frame forgery detection: A survey. Indian J Sci Technol 9(44)
89. Kingra S, Aggarwal N, Singh RD (2017) Inter-frame forgery detection in h. 264 videos using motion and brightness gradients. Multimed Tools Appl 76(24):25767–25786
90. Kobayashi K, Toda T (2018) Sprocket: Open-source voice conversion software. In: Odyssey, pp 203–210
91. Kobayashi M, Okabe T, Sato Y (2010) Detecting forgery from static-scene video based on inconsistency in noise level functions. IEEE Trans Inf Forensics Secur 5(4):883–892
92. Kohli A, Gupta A (2021) Detecting deepfake, faceswap and face2face facial forgeries using frequency cnn. Multimedia Tools and Applications, pp 1–18
93. Korshunov P, Halstead M, Castan D, Graciarena M, McLaren M, Burns B, Lawson A, Marcel S (2019) Tampered speaker inconsistency detection with phonetically aware audio-visual features. In: International conference on machine learning, CONF
94. Korshunov P, Marcel S (2018) Deepfakes:, a new threat to face recognition? assessment and detection. arXiv:1812.08685
95. Korshunov P, Marcel S (2018) Speaker inconsistency detection in tampered video. In: 2018 26Th european signal processing conference (EUSIPCO). IEEE, pp 2375–2379
96. Kumar A, Bhavsar A, Verma R (2020) Detecting deepfakes with metric learning. In: 2020 8Th international workshop on biometrics and forensics (IWBF). IEEE, pp 1–6
97. Kumar N, Kaur N, Gupta D (2020) Major convolutional neural networks in image classification: a survey. In: Proceedings of International Conference on IoT Inclusive Life (ICIIL 2019), NITTTR Chandigarh, India. Springer, pp 243–258
98. Kumar N, Kaur N, Gupta D (2020) Red green blue depth image classification using pre-trained deep convolutional neural network. Pattern Recognit Image Anal 30(3):382–390
99. Kumar P, Vatsa M, Singh R (2020) Detecting face2face facial reenactment in videos. arXiv:2001.07444
100. Laptev I, Marszałek M, Schmid C, Rozenfeld B (2008) Learning realistic human actions from movies
101. Lee S, Tariq S, Kim J, Woo S. S (2021) Tar:, Generalized forensic framework to detect deepfakes using weakly supervised learning. arXiv:2105.06117
102. Lee S, Yoo CD (2006) Video fingerprinting based on centroids of gradient orientations. In: 2006 IEEE International conference on acoustics speech and signal processing proceedings, vol 2. IEEE, pp II–II
103. Lee S, Yoo CD (2008) Robust video fingerprinting based on affine covariant regions. In: 2008 IEEE International conference on acoustics, speech and signal processing. IEEE, pp 1237–1240
104. Li H, Hu L, Wei L, Nagano K, Jaewoo S, Fursund J, Saito S Avatar digitization from a single image for real-time rendering (2020). US Patent 10,535,163
105. Li L, Bao J, Zhang T, Yang H, Chen D, Wen F, Guo B (2019) Face x-ray for more general face forgery detection. arXiv:1912.13458
106. Li M, Monga V (2012) Robust video hashing via multilinear subspace projections. IEEE Transactions on Image Processing 21(10):4397–4409
107. Li R, Liu Z, Zhang Y, Li Y, Fu Z (2018) Noise-level estimation based detection of motion-compensated frame interpolation in video sequences. Multimedia Tools and Applications 77(1):663–688

108. Li X, Lang Y, Chen Y, Mao X, He Y, Wang S, Xue H, Lu Q (2020) Sharp multiple instance learning for deepfake video detection. arXiv:2008.04585
109. Li Y, Chang M. C, Lyu S (2018) In ictu oculi:, Exposing ai generated fake face videos by detecting eye blinking. arXiv:1806.02877
110. Li Y, Yang X, Sun P, Qi H, Lyu S (2019) Celeb-df:, A new dataset for deepfake forensics. arXiv:1909.12962
111. Liu M, Ding Y, Xia M, Liu X, Ding E, Zuo W, Wen S (2019) Stgan: a unified selective transfer network for arbitrary image attribute editing. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3673–3682
112. Liu Y, Guan Q, Zhao X, Cao Y (2018) Image forgery localization based on multi-scale convolutional neural networks. In: Proceedings of the 6th ACM workshop on information hiding and multimedia security, pp 85–90
113. Liu Z, Luo P, Wang X, Tang X (2015) Deep learning face attributes in the wild. In: Proceedings of the IEEE international conference on computer vision, pp 3730–3738
114. Long C, Basharat A, Hoogs A (2019) A coarse-to-fine deep convolutional neural network framework for frame duplication detection and localization in forged videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 1–10
115. Lucas B. D, Kanade T et al (1981) An iterative image registration technique with an application to stereo vision
116. Malekesmaeili M, Fatourechi M, Ward RK (2009) Video copy detection using temporally informative representative images. In: 2009 International conference on machine learning and applications. IEEE, pp 69–74
117. Maras MH, Alexandrou A (2019) Determining authenticity of video evidence in the age of artificial intelligence and in the wake of deepfake videos. The Int J Evid Proof 23(3):255–262
118. Mase K (1991) Recognition of facial expression from optical flow. IEICE Trans Inf Syst 74(10):3474–3483
119. Masi I, Killekar A, Mascarenhas RM, Gurudatt S. P, AbdAlmageed W (2020) Two-branch recurrent network for isolating deepfakes in videos. arXiv:2008.03412
120. 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). IEEE, pp 83–92
121. Mehra A (2020) Deepfake detection using capsule networks with long short-term memory networks. Master's thesis, University of Twente
122. Milani S, Bestagini P, Tagliasacchi M, Tubaro S (2012) Multiple compression detection for video sequences. In: 2012 IEEE 14Th international workshop on multimedia signal processing (MMSP). IEEE, pp 112–117
123. Mirsky Y, Lee W (2021) The creation and detection of deepfakes: a survey. ACM Computing Surveys (CSUR) 54(1):1–41
124. Mittal T, Bhattacharya U, Chandra R, Bera A, Manocha D (2020) Emotions don't lie:, A deepfake detection method using audio-visual affective cues. arXiv:2003.06711
125. Mohammadi SH (2019) Text to speech synthesis using deep neural network with constant unit length spectrogram. US Patent 10,186,252
126. Montserrat DM, Hao H, Yarlagadda SK, Baireddy S, Shao R, Horváth J, Bartusiak E, Yang J, Güera D, Zhu F et al (2020) Deepfakes detection with automatic face weighting. arXiv:2004.12027
127. Nagothu D, Chen Y, Blasch E, Aved A, Zhu S (2019) Detecting malicious false frame injection attacks on surveillance systems at the edge using electrical network frequency signals. Sensors 19(11):2424
128. Nagothu D, Schwell J, Chen Y, Blasch E, Zhu S (2019) A study on smart online frame forging attacks against video surveillance system. In: Sensors and systems for space applications XII, vol 11017. International Society for Optics and Photonics, p 110170L
129. Nguyen HH, Fang F, Yamagishi J, Echizen I (2019) Multi-task learning for detecting and segmenting manipulated facial images and videos. arXiv:1906.06876
130. 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). IEEE, pp 2307–2311
131. Nguyen HM, Derakhshani R (2020) Eyebrow recognition for identifying deepfake videos. In: 2020 International conference of the biometrics special interest group (BIOSIG). IEEE, pp 1–5
132. Nguyen TT, Nguyen CM, Nguyen DT, Nguyen DT, Nahavandi S (2019) Deep learning for deepfakes creation and detection. arXiv:1909.11573

133. Nguyen XH, Tran TS, Nguyen KD, Truong DT et al (2021) Learning spatio-temporal features to detect manipulated facial videos created by the deepfake techniques. Forensic Science International: Digital Investigation 36:301108
134. Nirkin Y, Keller Y, Hassner T (2019) Fsgan: Subject agnostic face swapping and reenactment. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 7184–7193
135. Nirkin Y, Wolf L, Keller Y, Hassner T (2020) Deepfake detection based on the discrepancy between the face and its context. arXiv:2008.12262
136. Noguchi A, Yanai K (2010) A surf-based spatio-temporal feature for feature-fusion-based action recognition. In: European conference on computer vision. Springer, pp 153–167
137. Oord Avd, Dieleman S, Zen H, Simonyan K, Vinyals O, Graves A, Kalchbrenner N, Senior A, Kavukcuoglu K (2016) Wavenet:, A generative model for raw audio. arXiv:1609.03499
138. Oostveen J, Kalker T, Haitsma J (2002) Feature extraction and a database strategy for video fingerprinting. In: International conference on advances in visual information systems. Springer, pp 117–128
139. Ouyang J, Liu Y, Shu H (2017) Robust hashing for image authentication using sift feature and quaternion zernike moments. Multimed Tools Appl 76(2):2609–2626
140. Papadopoulou O, Zampoglou M, Papadopoulos S, Kompatsiaris Y, Teyssou D (2018) Invid fake video corpus v2. 0 (version 2.0) Dataset on Zenodo
141. Parkhi OM, Vedaldi A, Zisserman A (2015) Deep face recognition
142. Posters B (2018) Bill posters on instagram. artificially generated video of mark zuckerberg https://twitter.com/PressSec/status/1060374680991883265
143. Project A (2017) Ami corpus download. available at: http://groups.inf.ed.ac.uk/ami/download/
144. Project R Tools for digital forensics. http://www.rewindproject.eu/
145. Qadir G, Yahaya S, Ho AT (2012) Surrey university library for forensic analysis (sulfa) of video content
146. Qi H, Guo Q, Juefei-Xu F, Xie X, Ma L, Feng W, Liu Y, Zhao J (2020) Deeprhythm: exposing deepfakes with attentional visual heartbeat rhythms. In: Proceedings of the 28th ACM international conference on multimedia, pp 4318–4327
147. Rössler A, Cozzolino D, Verdoliva L, Riess C, Thies J, Nießner M (2018) Faceforensics: A large-scale video dataset for forgery detection in human faces. arXiv:1803.09179
148. Rössler A, Cozzolino D, Verdoliva L, Riess C, Thies J, Nießner M (2019) Faceforensics++: Learning to detect manipulated facial images. arXiv:1901.08971
149. Roy S, Sun Q (2007) Robust hash for detecting and localizing image tampering. In: 2007 IEEE International conference on image processing, vol 6. IEEE, pp VI–117
150. Sabir E, Cheng J, Jaiswal A, AbdAlmageed W, Masi I, Natarajan P (2019) Recurrent convolutional strategies for face manipulation detection in videos. Interfaces (GUI) 3:1
151. Saikia N (2015) Perceptual hashing in the 3d-dwt domain. In: 2015 International conference on green computing and internet of things (ICGCIot). IEEE, pp 694–698
152. Sanderson C (2019) Vidtimit audio-video dataset. available at: http://conradsanderson.id.au/vidtimit/
153. Saunders J, Comerford A, Williams G (2019) Detecting deep fakes with mice: Machines vs biology https://i.blackhat.com/USA-19/wednesday/us-19-williams-detecting-deep-Fakes-With-Mice-wp.pdf
154. Saxena S, Subramanyam A, Ravi H (2016) Video inpainting detection and localization using inconsistencies in optical flow. In: 2016 IEEE Region 10 conference (TENCON). IEEE, pp 1361–1365
155. Seeling P, Reisslein M (2001) Video traces research group http://trace.eas.asu.edu/
156. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision, pp 618–626
157. Shang Z, Xie H, Zha Z, Yu L, Li Y, Zhang Y (2021) Prrnet: Pixel-region relation network for face forgery detection. Pattern Recogn 116:107950
158. Shen J, Pang R, Weiss RJ, Schuster M, Jaitly N, Yang Z, Chen Z, Zhang Y, Wang Y, Skerrv-Ryan R et al (2018) Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In: 2018 IEEE International conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 4779–4783
159. Singh RD, Aggarwal N (2017) Detection of upscale-crop and splicing for digital video authentication. Digit Investig 21:31–52
160. Singh RD, Aggarwal N (2017) Optical flow and prediction residual based hybrid forensic system for inter-frame tampering detection. Journal of Circuits, Systems and Computers 26(07):1750107
161. Singh RD, Aggarwal N (2018) Video content authentication techniques: a comprehensive survey. Multimed Syst 24(2):211–240
162. Song F, Tan X, Liu X, Chen S (2014) Eyes closeness detection from still images with multi-scale histograms of principal oriented gradients. Pattern Recogn 47(9):2825–2838

163. Sowmya K, Chennamma H (2015) A survey on video forgery detection. Int J Comput Eng Appl 9(2):17–27
164. Stehouwer J, Dang H, Liu F, Liu X, Jain A (2019) On the detection of digital face manipulation. arXiv:1910.01717
165. Su Y, Xu J (2010) Detection of double-compression in mpeg-2 videos. In: 2010 2Nd international workshop on intelligent systems and applications. IEEE, pp 1–4
166. Sun K, Zhao Y, Jiang B, Cheng T, Xiao B, Liu D, Mu Y, Wang X, Liu W, Wang J (2019) High-resolution representations for labeling pixels and regions. arXiv:1904.04514
167. Sun Q, Liu Y, Chua T. S, Schiele B (2019) Meta-transfer learning for few-shot learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 403–412
168. Sun X, Wu B, Chen W (2020) Identifying invariant texture violation for robust deepfake detection. arXiv:2012.10580
169. Suwajanakorn S, Seitz SM, Kemelmacher-Shlizerman I (2017) Synthesizing obama: learning lip sync from audio. ACM Transactions on Graphics (TOG) 36(4):1–13
170. Tachibana H, Uenoyama K, Aihara S (2018) Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention. In: 2018 IEEE International conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 4784–4788
171. Tamgade SN, Bora VR (2009) Motion vector estimation of video image by pyramidal implementation of lucas kanade optical flow. In: 2009 Second international conference on emerging trends in engineering & technology. IEEE, pp 914–917
172. Tan M, Le QV (2019) Efficientnet:, Rethinking model scaling for convolutional neural networks. arXiv:1905.11946
173. Tariq S, Lee S, Woo SS (2020) A convolutional lstm based residual network for deepfake video detection. arXiv:2009.07480
174. Thies J, Elgharib M, Tewari A, Theobalt C (2019) Nießner, M.: Neural voice puppetry: Audio-driven facial reenactment. arXiv:1912.05566
175. Thies J, Zollhöfer M, Nießner M (2019) Deferred neural rendering: Image synthesis using neural textures. ACM Transactions on Graphics (TOG) 38(4):1–12
176. Thies J, Zollhofer M, Stamminger M, Theobalt C, Nießner M (2016) 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
177. Tian Y, Pei K, Jana S, Ray B (2018) Deeptest: Automated testing of deep-neural-network-driven autonomous cars. In: Proceedings of the 40th international conference on software engineering. ACM, pp 303–314
178. Todisco M, Wang X, Vestman V, Sahidullah M, Delgado H, Nautsch A, Yamagishi J, Evans N, Kinnunen T, Lee KA (2019) Asvspoof 2019:, Future horizons in spoofed and fake audio detection. arXiv:1904.05441
179. Tolosana R, Vera-Rodriguez R, Fierrez J, Morales A, Ortega-Garcia J (2020) Deepfakes and beyond:, A survey of face manipulation and fake detection. arXiv:2001.00179
180. TRECVID: Trec video retrieval evaluation. http://trecvid.nist.gov/
181. Tulyakov S, Liu MY, Yang X, Kautz J (2018) Mocogan: Decomposing motion and content for video generation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1526–1535
182. Verdoliva L (2020) Media forensics and deepfakes:, an overview. arXiv:2001.06564
183. Vincent J (2018) Jordan peele use ai to make barack obama deliver a psa about fake news https://www.theverge.com/tldr/2018/4/17/17247334/ai-fake-news-video%-barack-obama-jordan-peele-buzzfeed
184. Wahab AWA, Bagiwa MA, Idris MYI, Khan S, Razak Z, Ariffin MRK (2014) Passive video forgery detection techniques: a survey. In: 2014 10Th international conference on information assurance and security. IEEE, pp 29–34
185. Wan L, Wang Q, Papir A, Moreno IL (2018) Generalized end-to-end loss for speaker verification. In: 2018 IEEE International conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 4879–4883
186. Wang J, Wu Z, Chen J, Jiang YG (2021) M2tr:, Multi-modal multi-scale transformers for deepfake detection. arXiv:2104.09770
187. Wang Q, Li Z, Zhang Z, Ma Q (2014) Video inter-frame forgery identification based on optical flow consistency. Sensors & Transducers 166(3):229
188. Wang R, Juefei-Xu F, Huang Y, Guo Q, Xie X, Ma L, Liu Y (2020) Deepsonar:, Towards effective and robust detection of ai-synthesized fake voices. arXiv:2005.13770

189. Wang R, Juefei-Xu F, Ma L, Xie X, Huang Y, Wang J, Liu Y (2020) Fakespotter: a simple yet robust baseline for spotting ai-synthesized fake faces. In: International joint conference on artificial intelligence (IJCAI)
190. Wang SY, Wang O, Zhang R, Owens A, Efros AA (2020) Cnn-generated images are surprisingly easy to spot... for now. In: Proceedings of the IEEE conference on computer vision and pattern recognition, vol 7
191. Wang TC, Liu M. Y, Zhu J. Y, Liu G, Tao A, Kautz J, Catanzaro B (2018) Video-to-video synthesis. arXiv:1808.06601
192. Wang W, Farid H (2006) Exposing digital forgeries in video by detecting double mpeg compression. In: Proceedings of the 8th workshop on Multimedia and security. ACM, pp 37–47
193. Wang W, Farid H (2009) Exposing digital forgeries in video by detecting double quantization. In: Proceedings of the 11th ACM workshop on Multimedia and security. ACM, pp 39–48
194. Wang W, Jiang X, Wang S, Wan M, Sun T (2013) Identifying video forgery process using optical flow. In: International workshop on digital watermarking. Springer, pp 244–257
195. Wang Y, Skerry-Ryan R, Stanton D, Wu Y, Weiss RJ, Jaitly N, Yang Z, Xiao Y, Chen Z, Bengio S et al (2017) Tacotron:, Towards end-to-end speech synthesis. arXiv:1703.10135
196. Wheatley T, Weinberg A, Looser C, Moran T, Hajcak G (2011) Mind perception: Real but not artificial faces sustain neural activity beyond the n170/vpp PloS one 6(3)
197. Wiles O, Koepke A, Zisserman A (2018) Self-supervised learning of a facial attribute embedding from video. arXiv:1808.06882
198. Wodajo D, Atnafu S (2021) Deepfake video detection using convolutional vision transformer. arXiv:2102.11126
199. Xie W, Nagrani A, Chung JS, Zisserman A (2019) Utterance-level aggregation for speaker recognition in the wild. In: ICASSP 2019-2019 IEEE International conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 5791–5795
200. Xu F, Liu Y, Stoll C, Tompkin J, Bharaj G, Dai Q, Seidel HP, Kautz J, Theobalt C (2011) Video-based characters: creating new human performances from a multi-view video database. In: ACM SIGGRAPH 2011 Papers, pp 1–10
201. 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). IEEE, pp 8261–8265
202. Yoo DG, Kang SJ, Kim YH (2013) Direction-select motion estimation for motion-compensated frame rate up-conversion. J Disp Technol 9(10):840–850
203. Zampoglou M, Markatopoulou F, Mercier G, Touska D, Apostolidis E, Papadopoulos S, Cozien R, Patras I, Mezaris V, Kompatsiaris I (2019) Detecting tampered videos with multimedia forensics and deep learning. In: International conference on multimedia modeling. Springer, pp 374–386
204. Zhang X, Li H, Qi Y, Leow WK, Ng TK (2006) Rain removal in video by combining temporal and chromatic properties. In: 2006 IEEE International conference on multimedia and expo. IEEE, pp 461–464
205. Zhang Z, Robinson D, Tepper J (2018) Detecting hate speech on twitter using a convolution-gru based deep neural network. In: European semantic web conference, pp 745–760. Springer
206. Zhao T, Xu X, Xu M, Ding H, Xiong Y, Xia W (2020) Learning to recognize patch-wise consistency for deepfake detection. arXiv:2012.09311
207. Zhao Y, Wang S, Feng G, Tang Z (2010) A robust image hashing method based on zernike moments. J Comput Inf Syst 6(3):717–725
208. Zhu B, Fang H, Sui Y, Li L (2020) Deepfakes for medical video de-identification: Privacy protection and diagnostic information preservation. In: Proceedings of the AAAI/ACM conference on ai, ethics, and society, pp 414–420

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