Applied Intelligence
https://doi.org/10.1007/s10489-022-03766-z

Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward

Momina Masood[1] · Mariam Nawaz[1,2] · Khalid Mahmood Malik[3] · Ali Javed[2,3] · Aun Irtaza[4] · Hafiz Malik[4]

Accepted: 11 May 2022

Abstract
Easy access to audio-visual content on social media, combined with the availability of modern tools such as Tensorflow or Keras, and open-source trained models, along with economical computing infrastructure, and the rapid evolution of deep-learning (DL) methods have heralded a new and frightening trend. Particularly, the advent of easily available and ready to use Generative Adversarial Networks (GANs), have made it possible to generate deepfakes media partially or completely fabricated with the intent to deceive to disseminate disinformation and revenge porn, to perpetrate financial frauds and other hoaxes, and to disrupt government functioning. Existing surveys have mainly focused on the detection of deepfake images and videos; this paper provides a comprehensive review and detailed analysis of existing tools and machine learning (ML) based approaches for deepfake generation, and the methodologies used to detect such manipulations in both audio and video. For each category of deepfake, we discuss information related to manipulation approaches, current public datasets, and key standards for the evaluation of the performance of deepfake detection techniques, along with their results. Additionally, we also discuss open challenges and enumerate future directions to guide researchers on issues which need to be considered in order to improve the domains of both deepfake generation and detection. This work is expected to assist readers in understanding how deepfakes are created and detected, along with their current limitations and where future research may lead.

Keywords Artificial intelligence · Deepfakes · Deep learning · Face swap · Lip-synching · Puppetmaster · Speech synthesis · Voice conversion

1 Introduction

A boom in the availability of economical smart devices, such as cellphones, tablets, laptops, and digital cameras, has resulted in the exponential growth of digital multimedia content (e.g. images, audio, and video). Additionally, easy access to digital multimedia, along with the evolution of social media over the last decade, has allowed people to easily and rapidly share captured content. At the same time, we have witnessed tremendous advances in the field of machine learning (ML) with the introduction of sophisticated algorithms, like generative adversarial networks (GANs) [1], which can easily manipulate multimedia content and thus spread disinformation online through social media platforms. Moreover, today we live in a "post-truth" era, where a piece of information or disinformation may be utilized by malevolent actors to manipulate public opinion. Disinformation campaigns are very real, and have the potential to cause severe damage: election manipulation, defamation of any public person, or inflammation of popular sentiment. They may even be used to spark or justify a war. Given the ease with which false information may be created and spread it has become increasingly difficult to know what is true and trustworthy. One emerging technology is 'Deepfakes,' an AI-based synthesis or alteration of audio and visual content. The generation of deepfakes has advanced significantly, and they could be used to propagate

✉ Khalid Mahmood Malik
mahmood@oakland.edu

[1] Department of Computer Science, University of Engineering and Technology-Taxila, Taxila, Pakistan

[2] Department of Software Engineering, University of Engineering and Technology-Taxila, Taxila, Pakistan

[3] Department of Computer Science and Engineering, Oakland University, Rochester, MI, USA

[4] Electrical and Computer Engineering Department, University of Michigan-Dearborn, Dearborn, MI, USA

Published online: 04 June 2022

disinformation around the globe and may pose a severe threat, in the form of fake news, in the future [2], if they have not already.

Multimedia content as evidence is the current standard of proof in every sector of the legal world. It goes without saying that the audio-visual content admitted as evidence must be authentic and its integrity must be verified. At the same time, the introduction of easy to use manipulation tools (e.g. Zao [3], REFACE [4], FaceApp [5], Audacity [6], Soundforge [7]) has increased the perceived realism of fabricated data, which makes the authentication and integrity verification of such content even more challenging. Soon deepfakes are expected to be routinely used as weapons of disinformation, which will lead to a loss of credibility in state institutions, electronic media, and others due to the inability of common people to differentiate between original and fake videos. Moreover, the emergence of machine-generated text, along with manipulated audio-visual data, on social sites will bring more devastating effects and mislead decision-makers [8]. Currently, most of the existing multimedia forensic examiners focus on facing the challenge of analyzing multimedia files from social networks and sharing websites, e.g., YouTube, Facebook, etc. Satisfying the authentication and integrity requirements when flagging manipulated videos on social media is a challenging task because sophisticated deepfake generation algorithms with the potential to create more realistic fake videos have become more readily available.

Deepfake video can be categorized into the following types: i) face-swap ii) lip-synching iii) puppet-master iv) face synthesis and attribute manipulation, and v) audio-only deepfakes. In face-swap deepfakes, the face of the source person is replaced with the face of a victim to generate a fake video of the victim which in reality the source person has done. Face-swap-oriented deepfakes usually target a famous person by showing them in scenarios in which they never appeared in order to damage their reputation in the face of the public, for example, in non-consensual pornography. In lip-synching-based deepfakes, the movement of the target person's lips is manipulated to make them consistent with a specific audio recording so that the victim appears to say whatever is in the recording. In puppet-master deepfakes, video is created which mimics the expressions of the target person, such as eye movement, facial expressions, and head movement. Puppet-master deepfakes aim to hijack the source person's expression, or even full-body, in a video in order to animate it according to the impersonator's desire [9]. Face synthesis and attribute manipulation involve the generation of photo-realistic face images as well as facial attribute editing. This manipulation has been used to spread disinformation on social media using fake profiles. Lastly, audio deepfakes focus on the generation of the target speaker's voice using deep learning techniques to portray the speaker saying something they have not said [10, 11]. The fake voices can be generated using either text-to-speech synthesis (TTS) or voice conversion (VC). TTS aims to produce natural and intelligible voice waveforms, based on the provided text, that sounds like they have been spoken by the target identity. VC techniques transform the speech signal produced by a source speaker to seem like it was spoken by a target speaker while keeping the linguistic contents intact.

Unlike deepfake videos, less attention has been paid to the detection of audio deepfakes. In the last few years, voice manipulation has also become very sophisticated. Synthetic voices are not only a threat to automated speaker verification systems, but also to voice-controlled systems deployed in the Internet of Things (IoT) [12, 13]. Voice cloning has tremendous potential to destroy public trust and to empower criminals to manipulate business dealings, even private phone calls. For example, recently a case was reported in which bank robbers cloned a company executive's speech to dupe their subordinates into transferring hundreds of thousands of dollars into a secret account [14]. Voice cloning is expected to become a unique challenge in the future of deepfake detection. Therefore, it is important that unlike current approaches that focus only on detecting video signal manipulations, audio forgeries should also be examined.

Most of the existing surveys focus only on reviewing deepfake still images and video detection [15–17]. There is no recently published survey on deepfakes that specifically focuses on the generation and detection of both the audio and video. The discussion of generic image manipulation and multimedia forensic techniques was addressed in detail in [18], however deepfake generation techniques were not included. In [19], an overview of face manipulation and detection techniques was presented. Another survey, [20], reviewed visual deepfake detection approaches but does not discuss speech manipulation and its detection. The latest work presented by Mirsky et al. [21] gives an in-depth analysis of visual deepfake creation techniques,. Deepfake detection approaches are, however, only briefly discussed, and moreover, it lacks a discussion of audio deepfakes. To the best of our knowledge this paper is the first attempt to provide a detailed analysis and review of both audio and visual deepfake detection techniques and generative approaches. The following are the main contributions of our work:

i. To give the research community an insight into the various types of video and audio-based deepfake generation and detection methods.
ii. To provide the reader with the latest improvements, trends, limitations, and challenges in the field of audio-visual deepfakes.
iii. To give an understanding to the reader about the possible implications of audio-visual deepfakes.
iv. To act as a guide to the reader to understand the future trends of audio and visual deepfakes.

1.1 Literature collection and selection criteria

In this survey we reviewed the existing publications which approach techniques for the generation and detection of manipulated audio and video. A detailed description of the approach and protocols employed for the review is given in Table 1, Figs. 1 and 2.

The rest of the paper is organized as follows: Section 2 presents a discussion of deepfakes as a source of disinformation. In Section 3, the history and evolution of deepfakes are briefly discussed. Section 4 presents an overview of state-of-the-art audio and visual deepfake generation and detection techniques. Section 5 presents the details of available datasets used for both audio and video deepfakes detection. We have identified the open challenges for both audio-visual deepfake generation and detection in Section 6. In Section 7, we have discussed the possible future trends of both deepfake generation and detection, and finally, we conclude our work in Section 8.

2 Disinformation and misinformation using deepfakes

Misinformation is defined as false or inaccurate information that is communicated, regardless of an intention to deceive, whereas disinformation is the set of strategies employed to fabricate original "information" in order to achieve planned political or financial objectives, and is becoming increasingly prevalent. Because of the extensive use of social media platforms, it is now very easy to spread false information [22]. Although all categories of fake multimedia (i.e. video, images, and audio) could be sources of both disinformation and misinformation, audiovisual-based deepfakes are expected to be much more devastating. Historically, deepfakes were created to defame or discredit public figures. For example, in 2017 a female celebrity faced such situation when her fake pornographic video was circulated in cyberspace [20]. This is an evidence that deepfakes can be used to damage reputations, i.e., the character assassination of renowned people in order to defame them [20], blackmail of individuals for monetary benefits, or to create political or religious unrest by targeting politicians or religious figures with fake video/speech [23], etc. This damage is not limited to targeting individuals; rather deepfakes can be used to manipulate elections or even to theoretically start wars or used to deceive military analysts with fake information, and so on. Deepfakes are expected to advance these archtypes of disinformation and misinformation to the next level.

Trolls Trolls are hobbyists who spread inflammatory information solely to cause disorder or to get a reaction [14]. For

Table 1 Literature collection and preparation protocol

Preparation Protocol	Description
Purpose	• To provide a brief overview of existing state-of-the-art techniques and identify potential gaps in both audio-visual deepfake generation and detection. • To provide systematic review and structure to the existing state-of-the-art techniques with respect to each category of audio-visual deepfake generation and detection.
Data sources	Google Scholar, Springer Link, ACM digital library, IEEE explorer, and DBLP
Query	A methodical approach was designed to systematically utilize the data sources mentioned above and the following query strings were used: Deepfakes/Faceswap/ Face reenactment/ lip-syncing /Deepfakes AND Faceswap/ Deepfakes AND Face reenactment/ Deepfakes AND lip-syncing/ GAN synthesized/ face manipulation/ Attribute Manipulation/GAN AND Puppet Mastery/ GAN AND Expression Manipulation/ Video Synthesis/ Audio synthesis/ Deep learning AND TTS/ Deep learning AND Voice Conversion/ Deep learning AND Voice Cloning/Deepfakes AND Dataset/ Deepfakes AND Audio/ Deepfakes AND Video/Deepfakes AND image
Method	We have systematically categorized the literature on video and audio deepfakes as follows (Fig. 1): a) Video deepfake generation and detection into the following categories: face swap, lip-syncing, puppet-mastery, entire face synthesis, and facial attribute manipulation. b) Audio deepfake generation and detection into the following categories: text-to-speech synthesis and voice conversion.
Size	A total of 436 papers were retrieved using the method and query mentioned above from listed data sources up to −03-15-2022. We selected only those studies that were relevant and included the criteria 'deepfakes' in the positive set. Other relevant publications, where 'deepfake' was not in the positive set, were included in the negative set. All other studies were excluded from the final selection of papers, i.e., white papers and articles. The number of publications in the deepfake research area and distribution category, based on year, are presented in Fig. 2. It can be observed that the majority of the related articles are from conferences and informal publications i.e., arXiv. This is because the articles that have been published in top-tier journals make up only a small portion of the total currently available research.
Study types/inclusion and exclusion	Peer-reviewed journal papers and articles from conference proceedings were given more importance. Additionally, articles from archive literature were also considered.

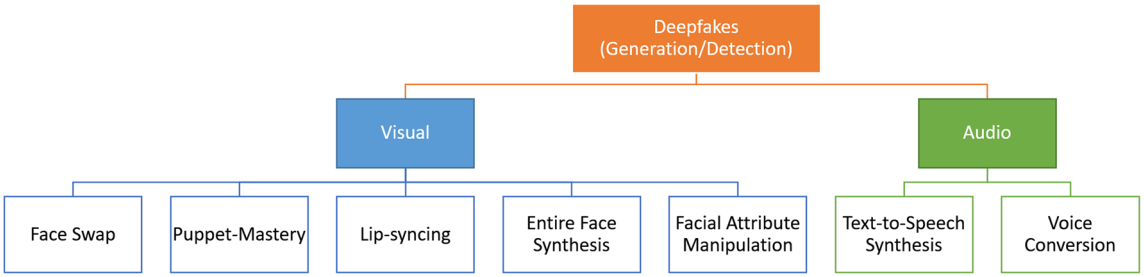

Fig. 1 Categorization of Audio and Visual Deepfakes

example, posting audio-visual manipulated racist or sexist content to promote hatred. Similarly, during the 2020 US presidential campaign, conflicting narratives about Trump and Biden were circulated on social media, contributing to an environment of fear [24]. In contrast to independent trolls, who spread the disinformation for their own satisfaction, hired trolls perform the same for monetary benefit. Different actors, like political parties, businessmen, and companies routinely hire people to forge news related to their competitors and spread it in the market [25]. Deepfake videos generated by hired trolls are the latest weapon in the ongoing fabricated news war that can bring a more devastating effect on society [26].

Bots Bots are automated software or algorithms used to spread fabricated or misleading content among the people [27]. A study published in [28] concluded that during the 2016 US presidential election, bots generated one-fifth of the tweets during the last month of the campaign. The emergence of deepfakes has bolstered the negative impact of bots i.e., recently, a messaging app named telegram used bots to post nude pictures of women [14].

Conspiracy theorists Conspiracy theorists range from nonprofessional filmmakers to Reddit agents who spread vague and doubtful claims on the internet either through "documentaries" or by posting stories and memes [29]. Recently, several conspiracy theorists have connected the current COVID pandemic with the China [30]. In such a situation, the use of fabricated audio-visual deepfake content by these theorists can increase controversy in global politics.

Hyper-partisan media Hyper-partisan media includes fake news websites and blogs which intentionally spread false information to a specific political demographic. Because of the extensive usage of social media, Hyper-partisan media is one of the biggest potential incubators for the spread of fabricated news [31]. Convincing AI-generated fake content assists these bloggers to easily spread disinformation, to attract visitors, or to increase views. As social platforms are largely independent and ad-driven mediums, spreading fabricated information may purely be a profit-making strategy [32].

Politicians One of the main sources of disinformation is the political parties themselves, which may spread manipulated information for point-scoring. Due to a large number of followers on social platforms, politicians are central nodes in online networks. So, politicians may use their fame and public support to spread false news among their followers. To defame opponent parties, politicians may use deepfakes to post controversial content about their competitors on conventional media [29].

Foreign governments As the Internet has converted the world into a "Global Village," it has become easier for conflicting countries to spread false news to target the reputation of any

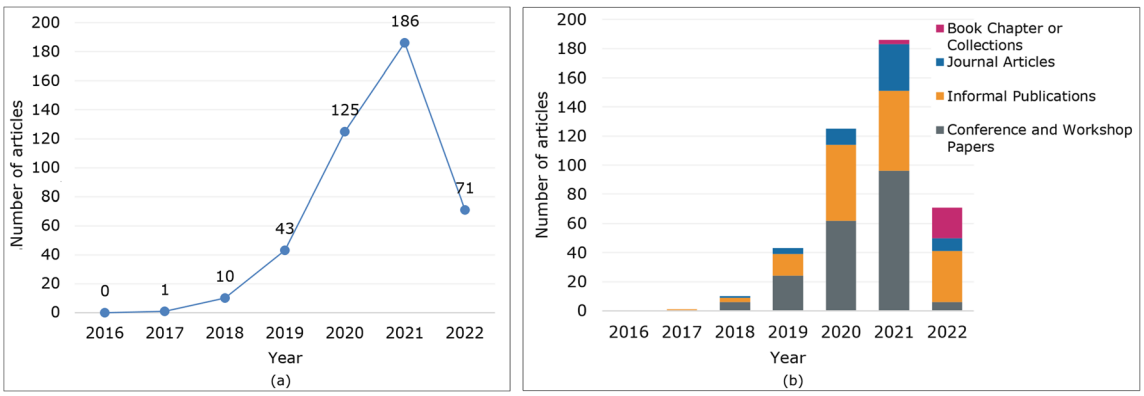

Fig. 2 Number of papers in the area of Deepfake research by **a** year-wise publication count, and **b** the number of publications by year belonging to studied categories, obtained from Google Scholar

country in the world. Many countries are running government-sponsored social media accounts, websites, and applications, contributing to political propaganda globally [14]. These non-state actors are anticipated to become more active in this sector as deepfakes techniques cut the costs of online propaganda. This raises the risk that extremist groups skilled in information warfare may exploit the technology and initiate foreign attacks on their own to increase the stress among countries.

3 DeepFakes evolution

The earliest example of manipulated multimedia content occurred in 1860 when a portrait of southern politician John Calhoun was skillfully manipulated by replacing his head with that of US President for propaganda purposes [33]. Usually, such manipulation is accomplished by adding (splicing), removing (inpainting), and replicating (copy-move) the objects within or between two images [18]. Then, suitable post-processing steps, such as scaling, rotating, and color adjustment are applied to improve the visual appearance, scale, and perspective coherence [34].

Aside from these traditional manipulation methods, advancements in Computer Graphics and deep learning (DL) techniques now offer a variety of different automated approaches for digital manipulation with better semantic consistency. A recent trend involves the synthesis of videos from scratch using autoencoders, or generative adversarial networks (GANs), for different applications [35] and, more specifically, photorealistic human face generation based on any attribute [36–39]. Another pervasive manipulation, called "shallow fakes" or "cheap fakes," are audio-visual manipulations created using cheaper and more accessible software. Shallow fakes involve basic editing of a video utilizing slowing, speeding, cutting, and selectively splicing together unaltered existing footage that can alter the whole context of the information delivered. In May 2019, a video of US Speaker Nancy Pelosi was selectively edited to make it appear that she was slurring her words and was drunk or confused [14]. The video was shared on Facebook and received more than 2.2 million views within 48 hours. Video manipulation for the entertainment industry, specifically in film production, has been done for decades. Figure 3 shows the evolution of deepfakes over the years. An early notable academic project was the Video Rewrite Program [40], intended for applications in movie dubbing, and published in 1997. It was the first software which was able to automatically reanimate facial movements in an existing video to a different audio track, and it achieved surprisingly convincing results.

The first true deepfake appeared online in September 2017 when a Reddit user named "deepfake" posted a series of computer-generated videos of famous actresses with their faces swapped onto pornographic content [20]. Another notorious deepfake case was the release of the deepNude application that allowed users to generate fake nude images [41]. This was the beginning of when deepfakes gained wider recognition within a large community. Today deepfake technology/applications, i.e. FakeApp [42], FaceSwap [43], and ZAO [3] are easily accessible, and users without a computer engineering background can create a fake video within seconds. Moreover, open-source projects on GitHub, such as DeepFaceLab [44] and related tutorials, are easily available on YouTube. A list of other available deepfake creation applications, software, and open-source projects is given in Table 2. Contemporary academic projects that lea to the development of deepfake technology are Face2Face [38] and Synthesizing Obama [37], published in 2016 and 2017

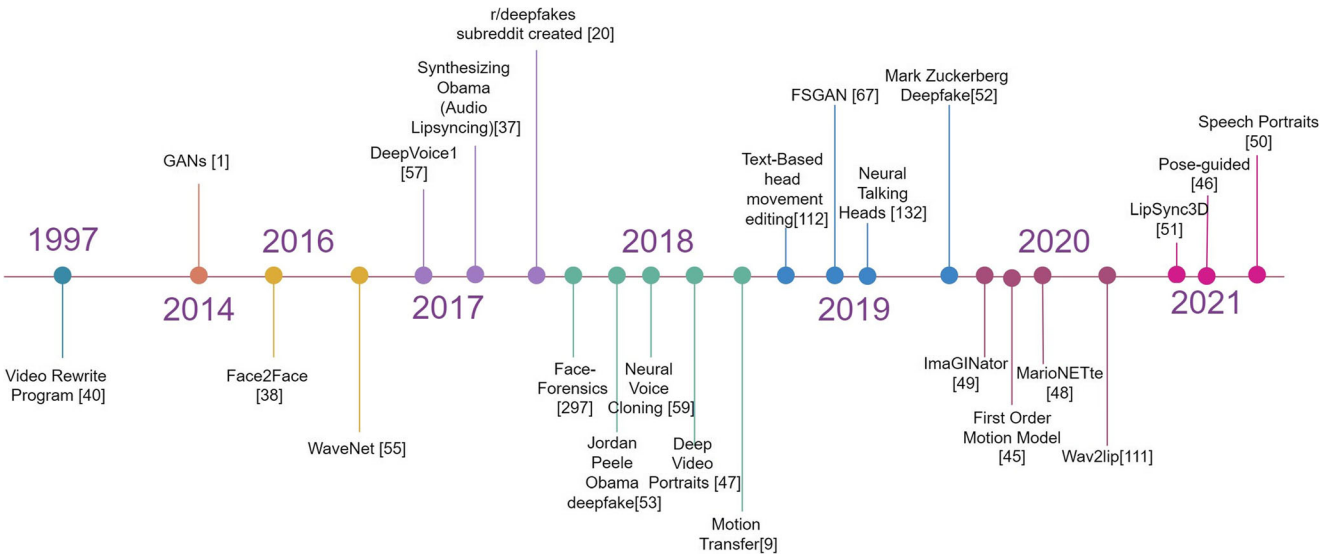

Fig. 3 Timeline of the evolution of Deepfakes

Table 2 An overview of Audio-visual deepfakes generation software, applications, and open-source projects

Tool	Type	Reference/Developer	Technique
Cheap fakes			
Adobe Premiere	Commercial Desktop Software	Adobe	Audio Video Editing, AI-powered video reframing
Corel VideoStudio	Commercial Desktop Software	Corel	Proprietary AI
Lip-sync			
dynalips	Commercial Web App	www.dynalips.com/	Proprietary
crazytalk	Commercial Web App	www.reallusion.com/crazytalk/	Proprietary
Wav2Lip	Open source implementation	github.com/Rudrabha/Wav2Lip	GAN with pre-trained discriminator network and visual quality loss function
Facial Attribute Manipulation			
FaceApp	MobileApp	FaceApp Inc	Deep generative CNNs
Adobe	Commercial Desktop Software	Adobe	DNNs + filters
Rosebud	Commercial Web App	www.rosebud.ai/	Proprietary AI
Face Swap			
ZAO	Mobile app	Momo Inc	Proprietary
REFACE	Mobile app	Neocortext, Inc	Proprietary
Reflect	Mobile app	Neocortext, Inc	Proprietary
Impressions	Mobile app	Synthesized Media, Inc.	Proprietary
FakeApp	Desktop App	www.malavida.com/en/soft/fakeapp/	GAN
FaceSwap	Open source implementation	faceswapweb.com/	Employed two seperate pairs of encoder-decoder with shared encoder parameters.
DFaker	Open source implementation	github.com/dfaker/df	-For facial reconstruction a DSSIM loss function is utilized. -Keras library-based implementation.
DeepFaceLab	Open source implementation	github.com/iperov/DeepFaceLab	- Several face extraction methods, e.g. dlib, MTCNN, S3FD etc. - Extends different Faceswap models i.e. H64, H128, LIAEF128, SAE [44].
FaceSwapGAN	Open source implementation	github.com/shaoanlu/faceswap-GAN	Uses two loss functions, namely adversarial loss and perceptual loss, to the auto-encoder.
DeepFake-tf	Open source implementation	github.com/StromWine/DeepFake-tf	Same as DFaker however, uses tensor-flow for implementation.
Faceswapweb	Commercial Web App	faceswapweb.com/	GAN
Face Reenactment			
Face2Face	Open source implementation	web.stanford.edu/~zollhoef/papers/CVPR2016_Face2Face/page.html	Uses 3DMM and ML technique
Dynamixyz	Commercial Desktop Software	www.dynamixyz.com/	Machine-learning
FaceIT3	Open source implementation	github.com/alew3/faceit_live3	GAN
Face Generation			
Generated Photos	Commercial Web App	generated.photos/	StyleGAN
Voice Synthesis			
Overdub	Commercial Web App	www.descript.com/overdub	Proprietary (AI based)
Respeecher	Commercial Web App	www.respeecher.com/	Combines traditional digital signal processing algorithms with proprietary deep generative modeling techniques
SV2TTS	Open source implementation	github.com/CorentinJ/Real-Time-Voice-Cloning	LSTM with Generalized end-to-end loss
ResembleAI	Commercial Web App	www.resemble.ai/	Proprietary (AI based)
Voicery	Commercial Web App	www.voicery.com/	Proprietary AI and deep learning
VoiceApp	Mobile app	Zoezi AB	Proprietary (AI-based)

respectively. Face2Face [38] captures the real-time facial expressions of the source person as they talk into a commodity webcam. It modifies the target person's face in the original video to depict them, mimicking the source facial expressions. Synthesizing Obama [37] is a video rewrite 2.0 program, used to modify mouth movements in video footage of a person in order to depict the person "saying" the words contained in an arbitrary audio clip. These works [37, 38] are focused on the manipulation of the head and facial region only. Recent development expands the application of deepfakes to the entire body [9, 45, 46], the generation of deepfakes from a single image [47–50], and temporally smooth video synthesis [51].

Most of the deepfakes currently present on social platforms like YouTube, Facebook or Twitter may be regarded as harmless, entertaining, or artistic. There are also some examples, however, where deepfakes have been used for revenge porn, hoaxes, political or non-political influence, and financial fraud [52]. In 2018, a deepfake video went viral online in which former U.S. President Barak Obama appeared to insult the current president, Donald Trump [53]. In June 2019, a fake video of Facebook CEO Mark Zuckerberg was posted to Instagram by the Israeli advertising company "Canny" [52]. More recently, extremely realistic deepfake videos of Tom Cruise posted on the TikTok platform gained 1.4million views within just a few days [54].

Apart from visual manipulation, audio deepfakes are a new form of cyber-attack, with the potential to cause severe damage to individuals due to highly sophisticated speech synthesis techniques i.e. WaveNet [55], Tacotron [56], and deep voice1 [57]. Fake audio-assisted financial scams increased significantly in 2019 as a direct result of the progression in speech synthesis technology. In August 2019, a European company's chief executive officer, tricked by an audio deepfake, made a fraudulent transfer of $243,000 [58]. A voice-mimicking AI software was used to clone the voice patterns of the victim by training ML algorithms using audio recordings obtained from the internet. If such techniques can be used to imitate the voice of a top government official or a military leader and applied at scale, it could have serious national security implications [59].

4 Audio-visual deepfake types and categorization of the literature

This section provides an in-depth analysis of existing state-of-the-art methods for audio and visual deepfakes. A review for each category of deepfake in terms of creation and detection is provided to give a deeper understanding of the various approaches. We provide a critical investigation of existing literature which includes the technologies, their capabilities, limitations, challenges, and future trends for both deepfake creation and detection. Deepfakes are broadly categorized into two groups, visual and audio manipulations, depending on the targeted forged modality (Fig. 1). Visual deepfakes are further grouped into the following types based on manipulation level: (i) face swap or identity swap, (ii) lip-syncing, (iii) face-reenactment or puppet-mastery, iv) entire face synthesis and v) facial attribute manipulation. Audio deepfakes are further classified as i) text-to-speech synthesis and ii) voice conversion.

Numerous models have been created to perform video manipulation. For manipulating both audio and video, different variants and combinations of GANs and encoder-decoder architectures are used. We have presented a generic pipeline for deepfakes generation in Fig. 4. To perform manipulation, an image or audio of the target identity and the conditioned source types including an image, video, sketch map, etc. are used. First, the facial region is detected and then cropped before translating both the target face and the source data to intermediate representations such as deep features, facial landmark keypoints, UV maps, and 3D morphable model parameters. The intermediate representations are then passed to different synthesis models, or combinations of models, such as GANs [1], encoder-decoder, Pix2Pix network [60], and RNN/LSTM. For audio deepfake generation the input can be either text or voice signal. In the case of text input, a linguistic analyzer is used to generate linguistic features such as phonemes, duration, and other different granularities. The obtained features are then passed to an acoustic analyzer for intermediate representation i.e., MCC (mel-cepstral coefficients), MGC (mel-generalized coefficients), and mel-spectrograms, etc., that are later used to generate output audio waveform. Finally, the output is acquired by re-rendering the generated face into the target frame. For the detection of audiovisual deepfakes, Fig. 5 shows general processing steps. Most of the deepfake detection approaches have employed either handcrafted features-based or deep learning-based methods for feature extraction. Few approaches are focused to employ the fusion of both handcrafted and deep features and using multiple modalities such as both audio and visual signals for effective manipulation. The computed key points are then used to classify the input media as real or fake. In the following sub-sections, we have analyzed the above-mentioned manipulation types in detail in terms of both synthesis and detection techniques.

4.1 Visual manipulations

4.1.1 Face-swap

Generation Visual manipulation is nothing new; images and videos have been forged since the early days of phototography. In face-swap [61], or face replacement, the face of the person in the source video is replaced by the face in the target video, as shown in Fig. 6. Traditional face-swap approaches [62–64] generally take three steps to perform a

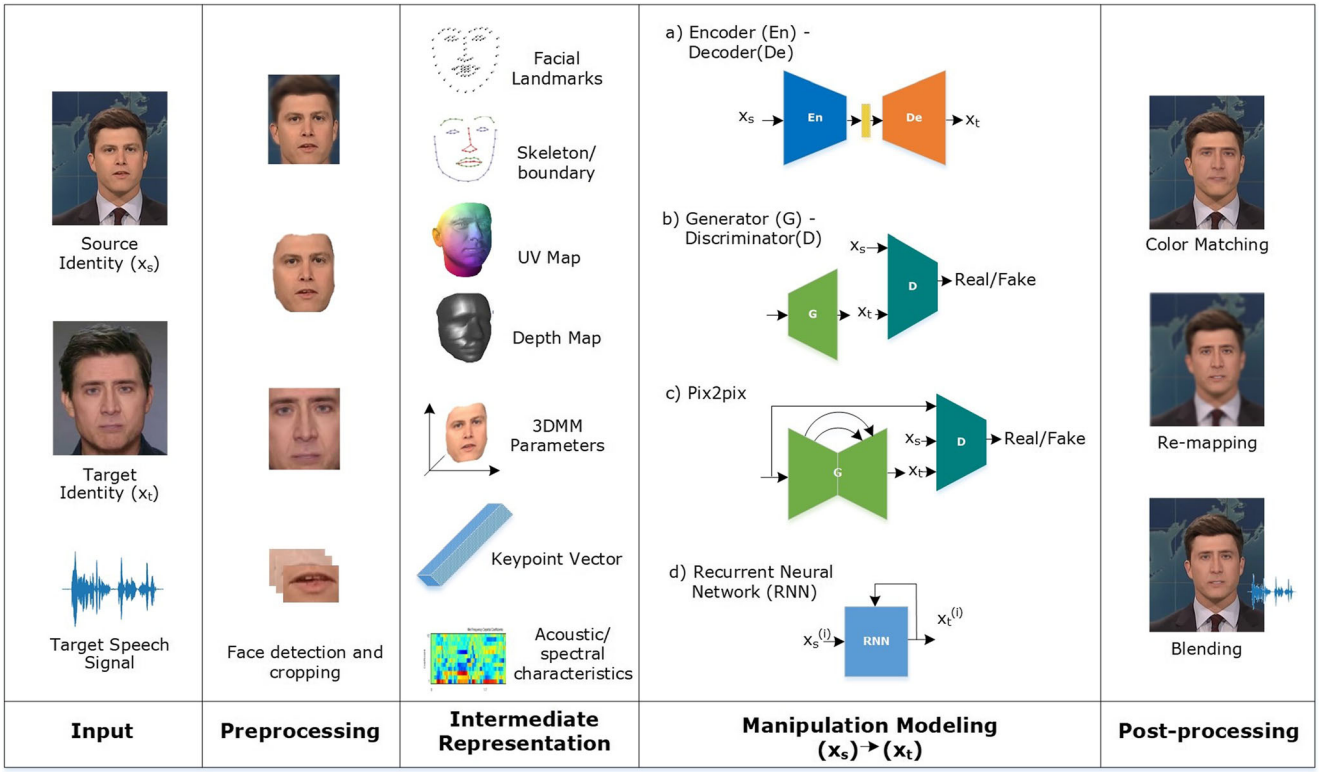

Fig. 4 Processing pipeline of audio-visual deepfakes generation approaches

face-swap operation. First, these tools detect the face in source images and then select a target's candidate face image from the facial library that is similar to the input facial appearance and pose. Second, the method replaces the eyes, nose, and mouth of the face and further adjusts the lighting and color of the candidate face image to match the appearance of input images, and seamlessly blends the two faces. Finally, the third step positions the blended candidate replacement by computing a match distance over the overlap region. These approaches generally offer good results but have two major limitations. First, they completely replace the input face with the target face, and expressions of the input face image are lost. Second, the synthetic result is very rigid, and the replaced face looks unnatural i.e., it requires a matching pose to generate good results.

Recently, DL-based approaches have become popular for synthetic media creation due to their realistic results. Recent deepfakes have shown how these approaches can be applied

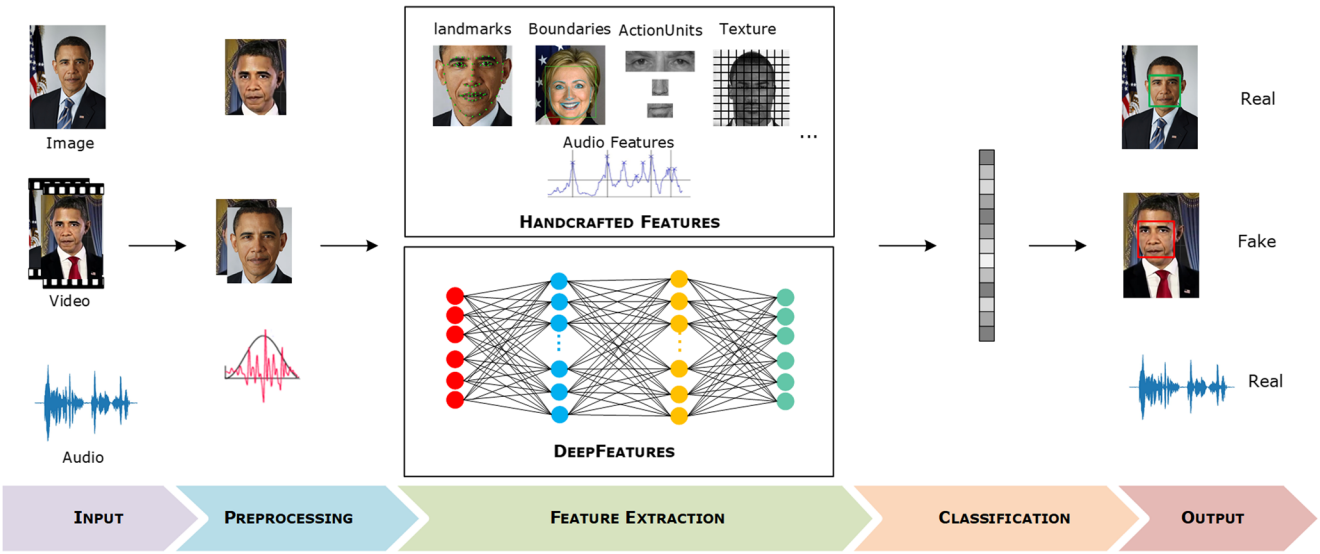

Fig. 5 The general processing pipeline for deepfake detection

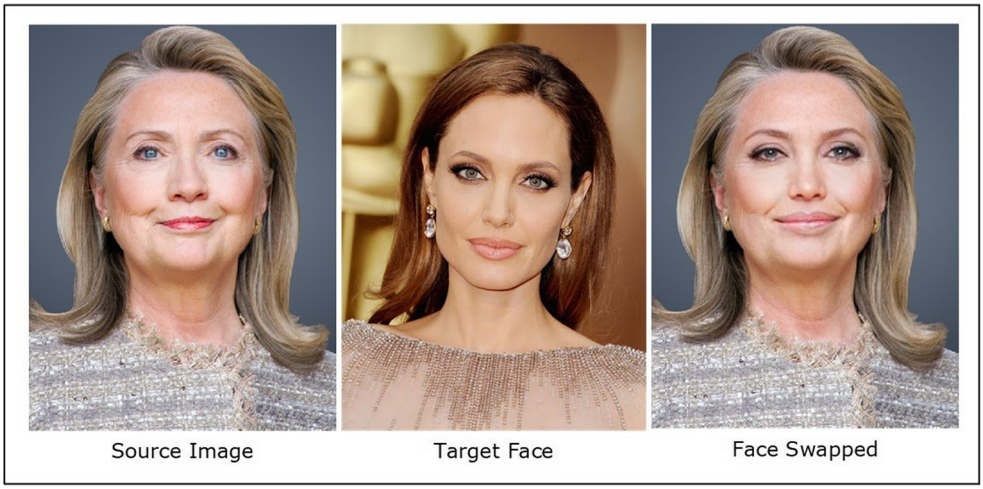

Fig. 6 A visual representation of Face-Swap based deepfakes

with automated digital multimedia manipulation. In 2017, the first deepfake video that appeared online was created using a face-swap approach, where the face of a celebrity was shown in pornographic content [20]. This approach used a neural network to morph a victim's face onto someone else's features while preserving the original facial expression. As time went on, face-swap software i.e. FakeApp [42] and FaceSwap [43] made it both easier and quicker to produce deepfakes with more convincing results by replacing the face in a video. These approaches typically use two encoder-decoder pairs. In this technique, an encoder is used to extract latent features of the face from the image and afterward the decoder is used to reconstruct the face. To swap faces between the source and target image, two pairs of encoder and decoder are required, where each encoder is first trained on the source and then the target image. Once training is complete, the decoders are swapped, so that an original encoder of the source image and a decoder of the target image are used to regenerate the target image with the features of the source image. The resulting image has the source's face on the target's face, while keeping the target's facial expressions. Fig. 7 is an example of a deepfake crafted in such a way that the feature set of face A is connected with the decoder B to reconstruct face B from the original face A. The recently launched ZAO [3], REFACE [4], and FakeApp [42] applications are more popular due to their effectiveness in producing realistic face swap-based deepfakes. FakeApp allows the selective modification of facial parts. ZAO and REFACE have gone viral lately, used by less tech-savvy users to swap their faces with movie stars and embed themselves into well-known movies and TV clips. There are many publicly available implementations of face-swap technology using deep neural networks, such as FaceSwap [43], DeepFaceLab [44], and FaceSwapGAN [65] leading to the creation of a growing number of synthesized media clips.

Until recently, most of the research focused on advances in face-swapping technology, either using a reconstructed 3D morphable model (3DMM) [61, 66], or GAN based models [65, 67]. Korshunova et al. [66] proposed a convolution neural network (CNN) based approach that transferred the semantic content, e.g., face posture, facial expression, and illumination conditions, of the input image to create the same effects in another image. They introduced a loss function that was a weighted combination of style loss, content loss, light loss, and total variation regularization. This method [66] generates more realistic deepfakes compared to [62], however, it requires a large amount of training data. Moreover, the trained model can be used to transform only one image at a time. Nirkin et al. [61] presented a method that used a full convolution network (FCN) for face segmentation and replacement in concert with a 3DMM to estimate facial geometry and corresponding texture. Then the face reconstruction was performed on the target image by adjusting the model parameters. These approaches [61, 66] have the limitation of subject-specific or pair-specific training. Recently subject agnostic approaches have been proposed to address this limitation [65, 67]. In [65], an improved deepfake generation approach using a GAN was proposed which adds adversarial loss and perceptual loss to VGGface, implemented in the auto-encoder architecture [43]. The addition of VGGFace perceptual loss made the direction of the eyes appear more realistic and consistent with the input, and also helped to smooth the artifacts added in the segmentation mask, resulting in a high-quality output video. FSGAN [67] allowed face swapping and reenactment in real-time by following the reenact and blend strategy. This method simultaneously manipulates pose, expression, and identity while producing high-quality and temporally coherent results. These GAN-based approaches [65, 67] outperform several existing autoencoder-decoder methods [42, 43] as they work without being explicitly trained on subject-specific images. Moreover, their iterative nature makes them well-suited for face manipulation tasks such as generating realistic images of fake faces.

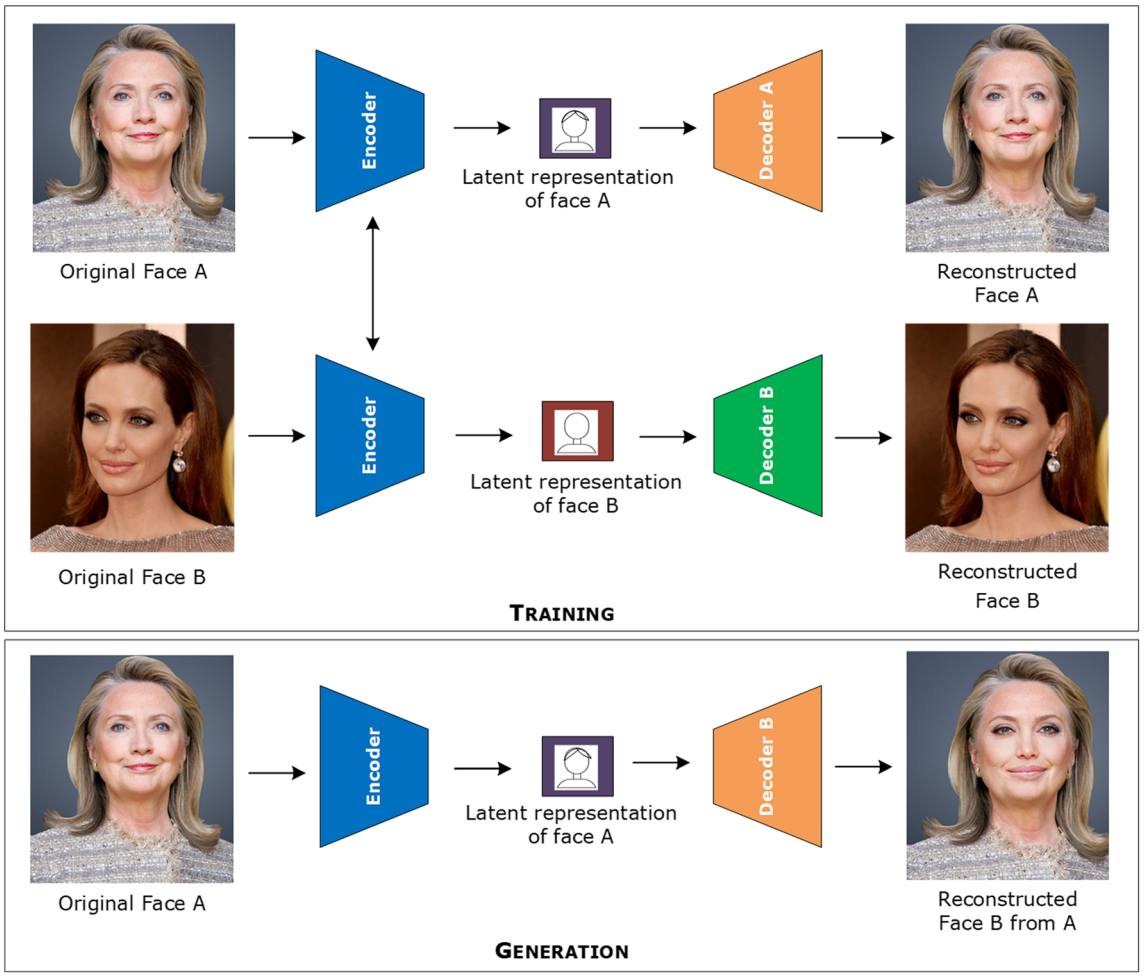

Fig. 7 Creation of a Deepfake using an auto-encoder and decoder. The same encoder-decoder pair is used to learn the latent features of the faces during training, while during generation decoders are swapped, such that latent face A is subjected to decoder B to generate face A with the features of face B

Some of the work used a disentanglement concept for face swap by using VAEs. RSGAN [68] employed two separate VAEs to encode the latent representation of facial and hair regions respectively. Both encoders were conditioned to predict the attributes that describe the target identity. Another approach, FSNet [69], presented a framework to achieve face-swapping using a latent space, to separately encode the face region of the source identity and landmarks of the target identity, which were later combined to generate the swapped face. However, these approaches [68, 69] do not preserve target attributes, like target occlusion and illumination conditions, well.

Facial occlusions are always challenging to handle in face-swapping methods. In many cases, the facial region in the source or target is partially covered with hair, glasses, a hand, or some other object. This results in visual artifacts and inconsistencies in the resultant image. FaceShifter [70] generates a swapped face with high fidelity and preserves the target attributes such as pose, expression, and occlusion. The identity encoder was used to encode the source identity and the target attributes, with feature maps being obtained via the U-Net decoder. These encoded features are passed to a novel generator with cascaded Adaptive Attentional Denormalization layers inside residual blocks which adaptively adjust the identity region and target attributes. Finally, another network is used to fix occlusion inconsistencies and refine the results. Table 3 presents details of Face-swap based deepfake creation approaches.

Detection Several recent studies have developed novel methods to identify face swap manipulations. Table 4, shows the comparison of faceswap detection techniques using both handcrafted and deep features.

Techniques based on handcrafted Features: Zhang et al. [73] propose a technique to detect swapped faces by using a Speeded Up Robust Features (SURF) descriptor for feature extraction. This is then used to train an SVM for classification, and then tested on a set of Gaussian blurred images. While this approach has improved deepfake image detection performance it is unable to detect manipulated videos. Yang et al. [74] introduce an approach to detect deepfakes by estimating the 3D head position from 2D facial landmarks. The

Table 3 An overview of Face-swap based deepfake generation techniques

Reference	Technique	Features	Dataset	Output Quality	Limitations
Faceswap [43]	Encoder-decoder	Facial landmarks	Private	256×256	▪ Blurry results due to lossy compression ▪ Lack of pose, facial expression, gaze direction, hairstyle, and lighting ▪ Requires massive number of target images
FaceSwapGAN [65]	GAN	VGGFace	VGGFace	256×256	▪ Lack of texture details; generates overly smooth results
DeepFaceLab [71]	Encoder-decoder	Facial landmarks	Private	256×256	▪ Fails to blend very different facial hues ▪ Requires target training data
Fast Face-swap [66]	CNN	VGGFace	▪ CelebA (200,000 images) ▪ Yale Face Database B (different pose and lighting conditions)	256×256	▪ Works for a single person only ▪ Gives better results for frontal face view ▪ Lack of skin texture details, e.g., smooth results and Facial Expression transfer ▪ Does not deal well with occluding objects i.e. glasses
Nirkin et al. [61]	FCN-8 s-VGG architecture	▪ Basel Face Model to represent faces ▪ 3DDFA model for expression	IARPA Janus CS2 (1275 face videos)	256×256	▪ Poor results in case of different image resolutions ▪ Fails to blend very different facial hues
Chen et al. [72]	VGG-16 net	68 facial landmarks	Helen (2330 images)	256×256	▪ Provide more realistic results but are sensitive to variation in posture and gaze
FSNet [69]	GAN	Facial landmarks	CelebA	128×128	▪ Sensitive to variation in angle
RSGAN [68]	GAN	Facial landmarks, segmentation mask	CelebA	128×128	▪ Sensitive to variation in angle, occlusion, lightning ▪ Limited output resolution
FaceShifter [70]	GAN	Attributes (face, occlusions, lighting or styles)	▪ VGG Face ▪ CelebA-HQ ▪ FFHQ	256×256	▪ Stripped artifacts

computed difference among the head poses is used as a feature vector to train an SVM classifier which is later used to differentiate between original and forged content. This technique exhibits good performance for deepfake detection but has a limitation in estimating landmark orientation in blurred images, which degrades the performance of this method under those conditions. Guera et al. [75] present a method for detecting synthesized faces from videos. Multimedia stream descriptors [76] are used to extract features, which are then used to train both an SVM and a random forest classifier to differentiate between the real and manipulated faces in a sample video. This technique gives an effective solution to deepfake detection but is unable to perform well against video re-encoding attacks. Ciftci et al. [77] introduce an approach to detect forensic changes within videos by computing biological signals (e.g. heart rate) from the face portion of the videos. Temporal and spatial characteristics of facial features are computed to train SVM and CNN models to differentiate between bonafide and fake videos. This technique has improved deepfake detection accuracy, however, it has a large feature vector space and its detection accuracy drops significantly when dimensionality reduction techniques are applied. Jung et al. [78] propose a technique to detect deepfakes by identifying an anomaly based on the time, repetition, and intervened eye-blinking duration within videos. This method combined the Fast-HyperFace [79] and EAR techniques (eye detect) [80] to detect eye blinking. An integrity authentication method is employed by tracking the fluctuation of eye blinks based on gender, age, behavior, and time factor to spot real and fake videos. The approach in [78] exhibits better deepfake detection performance, however, it is not appropriate if the subject in the video is suffering from mental illness, as abnormal eye blinking patterns are often observed in that population. Furthermore, the work in [81, 83] presents ML based approaches for face-swap detection, however, it still requires performance improvement in the presence of post-processing attacks.

Techniques based on Deep Features: Several studies have employed a DL-based method for Face-swap manipulation detection. Li et al. [84] proposed a method for detecting forensic modifications made within video. First, facial landmarks are extracted using the dlib software package [96].

Table 4 An overview of face swap deepfake detection techniques

Author	Technique	Features	Best Evaluation performance	Dataset	Limitations
Handcrafted features					
Zhang et al. [73]	SURF + SVM	64-D features using SURF	▪ Precision=97% ▪ Recall=88% ▪ Accuracy=92%	Generate deepfake dataset using LFW face database.	▪ Unable to preserve facial expressions ▪ Works with static images only.
Yang et al. [74]	SVM Classifier	68-D facial landmarks using DLib	ROC=89% ROC=84%	▪ UADFV ▪ DARPA MediFor GAN Image/ Video Challenge.	▪ Degraded performance for blurry images.
Guera et al. [75]	SVM, RF Classifier	Multimedia stream descriptor [76]	AUC=93% (SVM) AUC=96% (RF)	Custom dataset.	▪ Fails on video re-encoding attacks
Ciftci et al. [77]	CNN	medical signals features	Accuracy=96%	Face Forensics dataset	▪ Large feature vector space.
Jung et al. [78]	Fast-HyperFace [79], EAR [80]	Landmark features	Accuracy=87.5%	Eye Blinking Prediction dataset	▪ Inappropriate for people with mental illness
Matern et al. [81]	MLP, Logreg	16-D texture energy based features of eyes and teeth [82]	▪ AUC=.0.851(MLP) ▪ AUC=0.784 (LogReg)	FF++	▪ Only applicable to face images with open eyes and clear teeth.
Agarwal et al. [83]	SVM Classifier	16 AU's using OpenFace2 toolkit	AUC=93%	Own dataset.	▪ Degraded performance in cases where a person is looking off-camera.
Deep Learning-based features					
Li e al. [84]	VGG16, ResNet50, ResNet101, ResNet152	DLib facial landmarks	AUC=84.5 (VGG16), 97.4 (ResNet50), 95.4 (ResNet101), 93.8 (ResNet152)	DeepFake-TIMIT	▪ Not robust for multiple video compression.
Guera et al. [33]	CNN/ RNN	Deep features	Accuracy=97.1%	Customized dataset	▪ Applicable to short videos only (2 sec).
Li et al. [85]	CNN/RNN	DLib facial landmarks	TPR=99%	Customized dataset	▪ Fails over frequent closed eyes or blinking.
Montserrat et al. [86]	CNN+RNN	Deep features	Accuracy=92.61%	DFDC	▪ Performance needs improvement.
Lima et al. [87]	VGG11+LSTM	Deep features	Accuracy=98.26%, AUC= 99.73%	Celeb-DF	▪ Computationally complex.
Agarwal et al. [88]	VGG6+ encoder--decoder network	Deep features + behavioral biometrics	AUC=99% AUC=99% AUC=93% AUC=99%	WLDR FF DFD Celeb-DF	▪ Unable to generalize well to unseen deepfakes.
Fernandes et al. [89]	Neural-ODE model	Heart-rate	Loss=0.0215 Loss=0.0327	Custom DeepfakeTIMIT	▪ Computationally complex
Yang et al. [90]	GAN	Deep features	Accuracy=97.37% AUC=0.9999 AUC=0.9579 Accuracy=99.86%	FF++ CelebDF DFDC DeeperForensics	▪ Low generalization abilty
Sabir et al. [91]	CNN/RNN	Deep features	Accuracy=96.3%	FF++	▪ Results are reported for static images only.
Afchar et al. [92]	MesoInception-4	Deep features	TPR=81.3%	FF++	▪ Performance degrades on low quality videos.
Nguyen et al. [93]	CNN	Deep features	Accuracy=83.71%	FF++	▪ Degraded detection performance for unseen cases.
Stehouwer et al. [94]	CNN	Deep features	Accuracy=99.43%	Diverse Fake Face Dataset (DFFD)	▪ Computationally complex due to large feature vector space.
Rossle et al. [95]	SVM+CNN	Co-Occurance matrix + DF	Accuracy=90.29%	FF++	▪ Low performance on compressed videos.

Next, CNN-based models, named ResNet152, ResNet101, ResNet50, and VGG16 are trained to detect forged content from video. This approach is more robust in detecting forensic changes but it exhibits low performance on multi-time compressed videos. Guera et al. [33] propose a novel CNN to extract the features at the frame level. Then the RNN is trained on the set of extracted features to detect deepfakes from input video. This work achieves good detection performance but only on videos of short duration i.e. videos of 2 seconds or less. Li et al. [85] propose a technique to detect deepfakes by using the fact that the manipulated videos lack accurate eye blinking in synthesized faces. A CNN/RNN approach is used to detect a lack of eye blinking in the videos in order to expose the forged content. This technique shows better deepfake detection performance, however, it only uses the lack of eye blinking as a clue to detect the deepfakes. This approach has the following potential limitations: i) it is unable to detect forgeries in videos with frequent eye blinking, ii) it is unable to detect manipulated faces with closed eyes in training, and iii) it is inapplicable in scenarios where forgers can create realistic eye blinking in synthesized faces. Montserrat et al. [86] introduce a method for detecting visual manipulation in a video. Initially, a Multi-task convolutional neural network (MTCNN) [97] is employed to detect the faces from all video frames to compute the features. In the next step, an Automatic Face Weighting (AFW) mechanism, along with a Gated Recurrent Unit, is used to discard incorrectly identified faces. Finally, an RNN is employed to combine the features from all steps and locate the manipulated content in the video samples. The approach in [86] works well for deepfake detection, however, it is unable to obtain a prediction from the features in multiple frames. Lima et al. [87] introduce a technique to detect video manipulation by learning the temporal information of frames. Initially, VGG-11 is employed to compute the features from video frames, on which LSTM is applied for temporal sequence analysis. Several CNN frameworks, named R3D, ResNet, I3D, are trained on the temporal sequence descriptors outputted by the LSTM, in order to identify original and manipulated video. This approach [87] improves deepfake detection accuracy but at the expense of high computational cost. Agarwal et al. [88] present an approach to locate face-swap-based manipulations by combining both facial and behavioral biometrics. Behavioral biometrics are recognized with the encoder-decoder network (Facial Attributes-Net, FAb-Net) [98], whereas VGG-16 is employed for facial feature computation. Finally, by merging both metrics the inconsistencies in matching identities are revealed in order to locate face-swap deepfakes. The approach in [88] works well for unseen cases, however, it may not generalize well to lip-synch-based deepfakes. Fernandes et al. [89] introduce a technique to locate visual manipulation by measuring the heart-rate of the subjects. Initially, three techniques: skin color variation [99], average optical intensity [100], and Eulerian video magnification [101], are used to measure heart rate. The computed heart-rate was used to train a Neural Ordinary Differential Equations (Neural-ODE) model [102] to differentiate the original and altered content. This technique [89] works well for deepfake detection but also has increased computational complexity. In [103] a multi-scale texture difference network is introduced for face manipulation detection. The model is comprised of a ResNet-18 based textural difference information block and a multi-scale information extraction block. Then, the obtained features at different scales are fused to perform classification using cross-entropy loss. Yang et al. [90] propose a multi-scale self-texture attention deepfake detection framework based on facial texture analysis. The architecture work by identifying the potential texture difference between real and fake faces. It consists of a trace generator and a classification network. The trace generator network is comprised of an image analysis encoder followed by a self-texture attention module for the calculation of texture autocorrelation in features in order to differentiate between real and forged faces. For trace generation, the triplet loss is used to generate fake faces, and logistic regression for the actual face images. The loss function, based on probability-constrained trace control loss for trace construction and confined by classification probability, is applied. This method is robust to different textural post-processing operations, however the overall detection accuracy is low due to lack of generalizability. Other works [91–95] have explored CNN-based methods for the detection of swapped faces, however, there is a need for a more robust approach.

4.1.2 Lip-syncing

Generation The Lip-syncing approach involves synthesizing a video of a target identity such that the mouth region in the manipulated video is consistent with a specific audio input [37] (Fig. 8). A key aspect of the synthesis of a video with an audio segment is the movement and appearance of the lower portion of the mouth and its surrounding region. To convey a message more effectively and naturally, it is important to generate proper lip movements along with expressions. From a scientific point of view, lip-syncing has many applications in the entertainment industry, such as making audio-driven photorealistic digital characters in films or games, voice-bots, and dubbing films in foreign languages. Moreover, it can also help the hearing-impaired understand a scenario by lip-reading from a video created using genuine audio.

Existing works on lip-syncing [104, 105] require the reselection of frames from a video or transcription, along with target emotions, to synthesize lip motions. These approaches are limited to a dedicated emotional state and don't generalize well to unseen faces. However, DL models are capable of learning and predicting movements from audio features. A

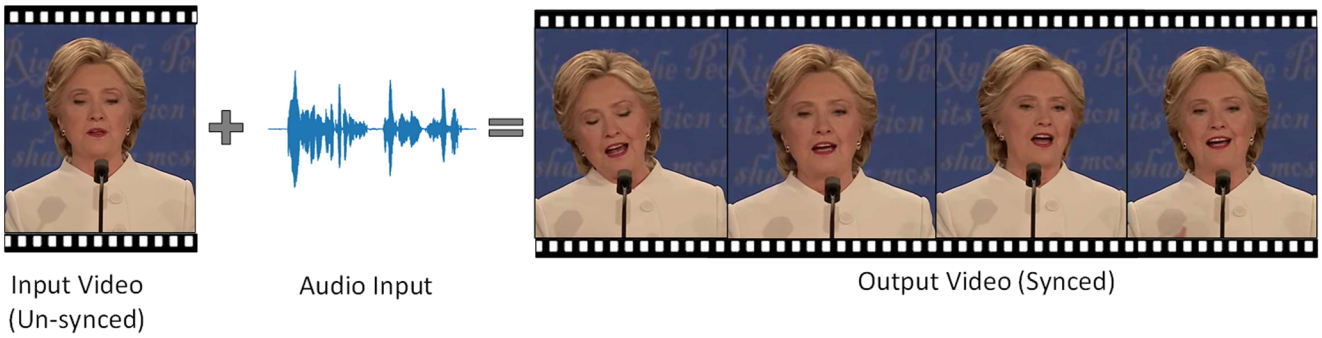

Fig. 8 A visual representation of lip-syncing of an existing video to an arbitrary audio clip

detailed analysis of existing DL-based methods used for Lip-sync-based deepfake generation are presented in Table 5. Suwajanakorn et al. [37] proposes an approach to generate a photo-realistic lip-synced video using a target's video and an arbitrary audio clip as input. A recurrent neural network (RNN) based model is employed to learn the mapping between audio features and mouth shape for every frame and later used frame reselection to fill in the texture around the mouth based on the landmarks. This synthesis is performed on the lower facial regions i.e. mouth, chin, nose, and cheeks, and applies a series of post-processing steps, such as smoothing jaw location and re-timing the video to align vocal pauses, or talking head motion, to produce videos that appear more natural and realistic. In this work, Barak Obama is considered as a case study due to the sufficient availability of online video footage. Thus, this model requires retraining and large amount of data for each individual. The Speech2Vid [106] model takes an audio clip and a static image of a target subject as input and generates a video that is lip-synced with the audio clip. This model uses Mel Frequency Cepstral Coefficient (MFCC) features, extracted from the audio input, and feeds them into a CNN-based encoder-decoder. As a post-processing step, a separate CNN is used for frame deblurring and sharpening in order to preserve the quality of visual content. This model generalizes well to unseen faces and thus does not need retraining for new identities. However, this work is unable to synthesize a variety of emotions on facial expressions.

Table 5 An overview of Lip sync-based deepfake generation techniques

Reference	Technique	Features	Dataset	Output Quality	Limitations
Suwajanakorn et al. [37]	RNN (single-layer unidirectional LSTM)	▪ Mouth landmarks (36-D features) ▪ MFCC audio features (28-D)	Youtube videos (17 hours)	2048×1024	▪ Requires a large amount of training data for the target person. ▪ Require retraining for each identity. ▪ Sensitive to the 3D movement of the head ▪ No direct control over facial expressions
Speech2Vid [106]	Encoder–decoder CNN	▪ VGG-M network ▪ MFCC audio features	▪ VGG Face ▪ LRS2 (41.3-hour video) ▪ VoxCeleb2 (test)	109×109	▪ lacks the ability to synthesize emotional facial expressions
Vougioukas et al. [107]	Temporal GAN	MFCC audio features	▪ GRID ▪ TCD TIMIT	96×128	▪ lacks the ability to synthesize emotional facial expressions flickering and jitter ▪ sensitive to large facial motions
Zhou et al. [108]	Temporal GAN	Deep audio-video features	▪ LRW ▪ MS-Celeb-1 M	256×256	▪ lacks the ability to synthesize emotional facial expressions
Vdub [109]	3DMM	▪ 66 facial feature points ▪ MFCC features	▪ Private	1024×1024	▪ Requires video of the target
LipGAN [110]	GAN	▪ VGG-M network ▪ MFCC features	▪ LRS 2	1280×720	▪ visual artifacts and temporal inconsistency ▪ unable to preserve source lip region characteristics
Wav2Lip [111]	GAN	▪ Mel-spectrogram representation	▪ LRS2	1280×720	▪ lacks the ability to synthesize emotional facial expressions

The GAN-based manipulations, such as [107] employ a temporal GAN, consisting of an RNN, to generate a photorealistic video directly from a still image and speech signal. The resulting video includes synchronized lip movements, eye-blinking, and natural facial expressions without relying on manually handcrafted audio-visual features. Multiple discriminators are employed to control frame quality, audio-visual synchronization, and overall video quality. This model can generate lip-syncing for any individual in real-time. In [108], an adversarial learning method is employed to learn disentangled audio-visual representation. The speech encoder is trained to project both audio and visual representations into the same latent space. The advantage of using a disentangled representation was that both the audio and video can serve as a source of speech information during the generation process. As a result, it is possible to generate realistic talking face sequences on an arbitrary identity with synchronized lip movement. Garrido et al. [109] present a Vdub system that captures the high-quality 3D facial model of both the source and the target actor. The computed facial model is used to photo-realistically reconstruct a 3D mouth model of the dubber to be applied on the target actor. An audio channel analysis is performed to better align the synthesized visual content with the audio. This approach better renders a coarse-textured teeth proxy, however it fails to synthesize a high-quality interior mouth region. In [110] a face-to-face translation method, LipGAN, is proposed which can synthesize a talking face video of any individual utilizing a given single image and audio segment as input. LipGAN consists of a generator network to synthesize portrait video frames with a modified mouth and jaw area from the given audio and target frames and uses a discriminator network to decide whether the synthesized face is synchronized with the given audio. This approach is unable to ensure temporal consistency in the synthesized content, as blurriness and jitter can be observed in the resultant video. Recently, Prajwal et al. [111] proposed a wav2lip speaker-independent model that can accurately synchronize lip movements in a video recording to a given audio clip. This approach employs a pre-trained lip-sync discriminator that is further trained on noisy generated videos in the absence of a generator. This model uses several consecutive frames instead of a single frame in the discriminator and employs visual quality loss, along with contrastive loss, thus increasing the visual quality by considering temporal correlation.

Recent approaches can synthesize photo-realistic fake videos from speech (audio-to-video) or text (text-to-video) with convincing video results. The methods proposed in [37, 112] can alter existing video of a person to the desired speech to be spoken from text input by modifying the mouth movement and speech accordingly. These approaches are more focused on synchronizing lip movements by synthesizing the region around the mouth. In [113] a VAE based framework is proposed to synthesize full pose video with facial expressions, gestures, and body posture movements from given audio.

Detection techniques based on handcrafted features Initially, ML-based methods are employed for the detection of lip-sync visual deepfakes. Korshunov et al. [114] propose a technique employing 40-D MFCC features containing the 13-D static, 13-D delta, and 13-D double-delta along with the energy, in combination with mouth landmarks to train four classifiers, i.e. SVM, LSTM, multilayer perceptron (MLP), and Gaussian mixture model (GMM). Three publicly available datasets, named VidTIMIT [115], AMI corpus [116], and GRID corpus [117] are used to evaluate the performance of this technique. From the results, it is concluded in [114] that the LSTM achieves better performance than other techniques. Lip-syncing deepfake detection performance of the LSTM method drops, however, for the VidTIMIT [115] and AMI [116] datasets due to fewer training samples for each person in both of these datasets over the GRID dataset. In [118] MFCC features were substituted with DNN embeddings i.e., language-specific phonetic features used for automatic speaker recognition. The evaluation show improved performance as compared to [114], however, performance is not evaluated on large-scale realistic datasets and GAN-based manipulation.

Techniques based on Deep Features: Other DL-based techniques, such as [119], propose a detection approach by exploiting the inconsistencies between phoneme-viseme pairs. In [119], the authors observe that in a video the lip shape associated with specific phenomes such as M, B, or P must be completely closed to pronounce them, however deepfake videos often lack this aspect. They analyze the performance by creating deepfakes using Audio-to-Video (A2V) [37] and Text-to-Video (T2V) [112] synthesis techniques. However, this method fails to generalize well for unseen samples during training. Haliassos et al. [120] propose a lip-sync deepfake detection approach, namely LipForensics, using a spatio-temporal network. Initially, a feature extractor 3D-CNN ResNet18 and a multiscale temporal convolutional network (MS-TCN) are trained on lip-reading datasets such as Lipreading in the Wild (LRW). Then, the model is fine-tuned on deepfake videos using the FaceForensics++ (FF++) dataset. This method also performs well over different post-processing operations such as blur, noise, compression etc., however, the performance substantially decreases when there is a limited mouth movement in the video, such as pauses in speech or less movement in the lips. Chugh et al. [121] propose a deepfake detection mechanism by finding a lack of synchronization between the audio and visual channels. They compute a modality dissimilarity score (MDS) between the audio and visual modalities. A sub-network based on 3D-ResNet architecture is used for feature computation and employs two loss functions, a cross-entropy loss at the output

layer for robust feature learning, and a contrastive loss, computed over the segment-level audiovisual features. The MDS is calculated as the total audiovisual dissonance over all segments of the video and is used for the classification of the video as real or fake. Mittal et al. [122] proposes a Siamese network architecture for audio-visual deepfake detection. This approach compares the correlation between emotion-based differences in facial movements and speech in order to distinguish between real and fake. However, this approach requires a real-fake video pair for the training of the network and fails to classify correctly if only a few frames in the video have been manipulated. Chintha et al. [123] propose a framework based on the XceptionNet CNN for facial feature extraction and then pass it to a bidirectional LSTM network for the detection of temporal inconsistencies. The network is trained via two loss functions, i.e., cross-entropy and KL-divergence, to discriminate the feature distribution of real video from that of manipulated video. Table 6 presents a comparison of handcrafted and deep learning techniques employed for the detection of lip sync-based deepfakes.

4.1.3 Puppet-master

Generation Puppet-master, also known as face reenactment, is another common variety of deepfake that manipulates the facial expressions of a person, e.g., transferring the facial gestures, eye, and head movements, to an output video that reflect those of the source actor [124] as shown in Fig. 9. Puppet-mastery aims to deform the person's mouth movement to make fabricated content. Facial reenactment has various applications, like altering the facial expression and mouth movement of a participant to a foreign language in an online multilingual video conference, dubbing or editing an actor's head and their facial expressions in film industry post-production systems, or creating photorealistic animation for movies and games, etc.

Initially, 3D facial modeling-based approaches for facial reenactment were proposed because of their ability to accurately capture the geometry and movement, and for improved photorealism in reenacted faces. Thies et al. [125, 126] presented the first real-time facial expression transfer method from an actor to a target person. A commodity RGB-D sensor was used to track and reconstruct the 3D model of the source and target actors. For each frame, the tracked deformations of the source face were applied to the target face model, and later the altered face was blended onto the original target face while preserving the facial appearance of the target face model. Face2Face [38] is an advanced form of facial reenactment technique, as presented in [125]. This method worked in real-time and was capable of altering the facial movements of generic RGB video streams, e.g., YouTube videos, using a standard webcam. The 3D model reconstruction approach was combined with image rendering techniques to generate the output. This could create a convincing and instantaneous re-rendering of a target actor with a relatively simple home setup. This work was further extended to control the facial expressions of a person in a target video based on intuitive hand gestures using an inertial measurement unit [127].

Later, GANs were successfully applied for facial reenactment due to their ability to generate photo-realistic images. Pix2pixHD [60] produced high-resolution images with better fidelity by combining a multi-scale conditional GANs (cGAN) architecture [128] using a perceptual loss. Kim et al. [47] proposed an approach that allowed the full

Table 6 An overview of Lip sync-based deepfake detection techniques

Author	Technique	Performance reported	Dataset used	Limitations
Handcrafted features				
Korshunov et al. [114]	SVM, LSTM, MLP, GMM	EER=24.74 (LSTM), 53.45 (MLP), 56.18(SVM), 56.09(GMM) EER=33.86 (LSTM), 41.21(MLP), 48.39(SVM), 47.84 (GMM) EER=14.12 (LSTM), 28.58(MLP), 30.06 (SVM), 46.81(GMM)	VidTIMIT AMI GRID	LSTM performs better than others but its performance degrades as the training samples decrease.
Agarwal et al. [119]	SVM	Accuracy=99.6%	Custom dataset	Performance degrades for unseen samples
Deep Learning-based features				
Haliassos et al. [120]	3D-ResNet18, multi-scale temporal convolutional network	AUC=97.1%	FF++	Performance degrades in cases when there is limited lip movement
Mittal et al. [122]	siamese network architecture	Accuracy =84.4% AUC=96.3%(LQ), 94.9%(HQ)	DFDC DF-TIMIT	Requires a real–fake video pair for training.
Chintha et al. [123]	XceptionNet CNN with bidirectional LSTM network	Accuracy=97.83% Accuracy=96.89%	Celeb-Df FF++	Performance degrades on compressed samples

Fig. 9 A visual representation of puppet-master based deepfake

reanimation of portrait videos by an actor, such as changing head pose, eye gaze, and blinking, rather than just modifying the facial expression of the target identity and thus produced photorealistic dubbing results. At first, a face reconstruction approach was used to obtain a parametric representation of the face and illumination information from each video frame to produce a synthetic rendering of the target identity. This representation was then fed to a render-to-video translation network based on the cGAN in order to predict the synthetic rendering into photo-realistic video frames. This approach required training the videos for target identity. Wu et al. [129] proposed ReenactGAN, which encodes input facial features into a boundary latent space. A target-specific transformer was used to adapt the source boundary space according to the specified target, and later the latent space was decoded onto the target face. GANimation [130] employed a dual cGAN generator conditioned on emotion action units (AU) to transfer facial expressions. The AU-based generator used an attention map to interpolate between the reenacted and original images. Instead of relying on AU estimations, GANnotation [131] used facial landmarks, along with a self-attention mechanism, for facial reenactment. This approach introduced a triple consistency loss to minimize visual artifacts but required the images to be synthesized with a frontal facial view for further processing. These models [130, 131] required a large amount of training data for the target identity to perform well at oblique angles or they lacked the ability to generate photo-realistic reenactment for unknown identities.

Recently, few-shot or one-shot face reenactment approaches have been proposed to achieve reenactment using a few, or even a single, source image. In [39], a self-supervised learning model, X2face, using multiple modalities such as driving frame, facial landmarks, or audio, to transfer the pose and expression of the input source to the target expression, was proposed. X2face uses two encoder-decoder networks: an embedding network and a driving network. The embedding network learns face representation from the source frame and the driving network learns pose and expression information from the driving fame to the vector map. The driving network was crafted to interpolate face representation from the embedded network in order to produce target expressions. Zakharov et al. [132] present a meta-transfer learning approach where the network was first trained on multiple identities and then fine-tuned on the target identity. First, target identity encoding is obtained by averaging the target's expressions and associated landmarks from different frames. Then a pix2pixHD [60] GAN was used to generate the target identity using source landmarks as input, and identity encoding via adaptive instance normalization (AdaIN) layers. This approach works well at oblique angles and directly transfers the expression without requiring intermediate boundary latent space or an interpolation map, as in [39]. Zhang et al. [133] propose an auto-encoder-based structure to learn the latent representation of the target's facial appearance and the source's face shape. These features are used as input to SPADE residual blocks for the face reenactment task, which preserves the spatial information and concatenates the feature map in a multi-scale manner from the face reconstruction decoder. This approach can better handle large pose changes and exaggerated facial actions. In FaR-GAN [134], learnable features from convolution layers are used as input to the SPADE module instead of using multi-scale landmark masks, as in [133]. Usually, few-shot learning fails to completely preserve the source identity in the generated results for cases where there is a large pose difference between the reference and target image. MarioNETte [48] is proposed to mitigate identity leakage by employing attention block and target feature alignment. This helps the model to accommodate the variations between face structures better. Finally, the identity is retained by using a novel landmark transformer, influenced by the 3DMM facial model [135].

Real-time face reenactment approaches, such as FSGAN [67], perform both facial replacement and reenactment with occlusion handling. For reenactment, a pix2pixHD [60]

generator takes the target's image and source's 3D facial landmark as input and outputs a reenacted image and 3-channel (hair, face, and background) encoded segmentation mask. The recurrent generator is trained recursively where output is iterated multiple times for incremental interpolation from source to target landmarks. The results are further improved by applying Delaunay Triangulation and barycentric coordinate interpolation to generate output similar to the target's pose. This method achieves real-time facial reenactment at 30fps and can be applied to any face without requiring identity-specific training. Table 7 provides the summary of techniques adopted for facial expression manipulation mentioned above.

In the next few years, photo-realistic full-body reenactment [9, 136] videos will also be viable, where the target's expression, along with mannerisms, will be manipulated to create realistic deepfakes. The videos that will be generated using the above-mentioned techniques will be further merged with fake audio to create completely fabricated content [137]. These progressions enable the real-time manipulation of facial expressions and motion in videos while making it challenging to distinguish between what is real and what is fake.

Detection Techniques based on handcrafted Features: Matern et al. [81] presented an approach for classifying forged content by employing simple facial handcrafted features like the color of eyes, missing artifact information in the eyes and teeth, and missing reflections. These features were used to train two models, i.e. logistic regression and MLP, to distinguish manipulated content from the original data. This technique has a low computational cost; however, it applied only

Table 7 An overview of face reenactment-based deepfake generation techniques

Reference	Technique	Features	Dataset	Output Quality	Limitations
Face2Face [38]	3DMM	▪ parametric model ▪ Facial landmark features	customized	1024×1024	▪ Sensitive to facial occlusions
Kim et al. [47]	cGAN	parametric model of the face (261 parameters/frame)	customized	1024×1024	▪ 1–3 min. Video of target ▪ Sensitive to facial occlusions
ReenactGAN [129]	GAN	Facial landmark features	▪ CelebV dataset ▪ WFLW Dataset ▪ Helen, DISFA	256×256	▪ 30 min. Video of target ▪ Lacks gaze adaptation
GANimation [130]	GAN (2 Encoder- 2 Decoder)	AUs	▪ EmotioNet dataset ▪ RaFD dataset	128×128	▪ Lack of pose and gaze adaptation
GANnotation [131]	GAN	Facial landmark features	▪ 300-VWChallenge dataset ▪ BP4D dataset ▪ Helen, LFPW, AFW, IBUG, and a subset of multiple datasets	128×128	▪ Lack of gaze adaptation
X2face [39]	2Encoder-2Decoder	▪ Facial landmark features ▪ 256-D audio features	▪ VGG Face dataset ▪ VoxCeleb dataset ▪ AFLW dataset	256×256	▪ Wrinkle artifacts ▪ Lack of gaze adaptation
Zakharov et al. [132]	GAN (1Encoder-2Decoder)	Facial landmark features	VoxCeleb dataset	256×256	▪ Sensitive to source identity leakage ▪ Lack of gaze adaptation
Zhang et al. [133]	GAN (1Encoder-2Decoder)	Appearance and shape feature Map	▪ VGG Face dataset ▪ WFLW ▪ EOTT dataset ▪ CelebA-HQ dataset ▪ LRW dataset.	256×256	▪ Low visual quality output (256×256)
FaR-GAN [134]	GAN	Facial landmark and Boundary features	▪ VGG Face dataset ▪ VoxCeleb1 dataset	256×256	▪ Sensitive to source identity leakage ▪ Lack of gaze adaptation
MarioNETte [48]	GAN (2Encoder-1Decoder)	Facial landmark features	▪ VoxCeleb1	256×256	▪ Fails to preserve source facial characteristics completely
FSGAN [67]	GAN+RNN	▪ Facial landmarks ▪ LFW parts label set	▪ IJB-C dataset (5500 face videos) ▪ VGGFace2 ▪ CelebA ▪ Figaro dataset	256×256	▪ The identity and texture quality degrade in case of large angular differences ▪ Fail to fully capture facial expressions ▪ blurriness in image texture ▪ limited to the resolution of training data

to the visual content with open eyes or visible teeth. Amerini et al. [138] proposed an approach based on optical flow fields to detect synthesized faces in digital videos. The optical flow fields [139] of each video frame were computed using PWC-Net [140]. The estimated optical flow fields of frames were used to train the VGG16 and ResNet50 to classify real and fake content. This method [138] exhibited better deepfake detection performance, however, only initial results have been reported. Agarwal et al. [83] presented a user-specific technique for deepfake detection. First, a GAN was used to generate all three types of deepfakes for US ex-president Barack Obama. Then the OpenFace2 [141] toolkit was used to estimate facial and head movements. The estimated difference between the 2D and 3D facial and head landmarks was used to train a binary SVM to classify between the original face and synthesized face of Barack Obama. This technique provided good detection accuracy, however, it was vulnerable in those scenarios where a person is looking off-camera.

Techniques based on Deep Features: Several research works have focused on employing DL-based methods for puppet-mastery deepfake detection. Sabir et al. [91] observed that while generating the manipulated content, forgers often do not impose temporal coherence in the synthesis process. So, in [91], a recurrent convolutional model was used to investigate the temporal artifacts in order to identify synthesized faces in the images. This technique [91] achieved better detection performance, however, it worked best on static frames. Rossler et al. [95] employed both handcrafted (co-occurrence matrix) and learned features for detecting manipulated content. It was concluded in [95] that the detection performance of both networks, either employing hand-crafted or deep features, degraded when evaluating them on compressed videos. To analyze the mesoscopic properties of manipulated content, Afchar et al. [92] proposed an approach where they employed two variants of the CNN model with a small number of layers, named Meso-4 and MesoInception-4. This method managed to reduce the computational cost by downsampling the frames but at the expense of a decrease in accuracy in deepfake detection. Nguyen et al. [93] proposed a multi-task, learning-based CNN network to simultaneously detect and localize manipulated content from videos. An autoencoder was used for the classification of forged content, while a y-shaped decoder was applied to share the extracted information for the segmentation and reconstruction steps. This model was robust to deepfake detection; however, the evaluation accuracy degraded when presented with unseen scenarios. To overcome the issue of performance degradation, as in [93], Stehouwer et al. [94] proposed a Forensic transfer (FT) based CNN approach for deepfake detection. This work [94], however, suffered from high computational cost due to a large feature space. The comparison of these handcrafted and deep features-based face reenactment deepfake detection techniques is presented in Table 8.

4.1.4 Face synthesis

Generation Facial editing in digital images has been heavily explored for decades. It has been widely adopted in the art, animation, and entertainment industry, however lately it has been exploited to create deepfakes for identity impersonation. Face generation involves the synthesis of photorealistic images of a human face that may or may not exist in real life. The tremendous evolution in deep generative models has made them widely adopted tools for face image synthesis and editing. Generative deep learning models, i.e. GAN [1] and VAE [142], have been successfully used to generate photo-realistic fake human face images. In facial synthesis, the objective is to generate non-existent but realistic-looking faces. Face synthesis has enabled a wide range of beneficial applications, like automatic character creation for video games and 3D face modeling industries. AI-based face synthesis could also be used for malicious purposes such as the synthesis of photorealistic fake profile picture for a fake social network account in order to spread disinformation. Several approaches have been proposed to generate realistic-looking facial images that humans are unable to recognize as synthesized. Figure 10 shows the improvement in the quality of synthetic facial images between 2014 and 2019. Table 9 provides a summary of works presented for the generation of entirely synthetic faces.

Since the emergence of GANs [1] in 2014, significant efforts have been made to improve the quality of synthesized images. The images generated using the first GAN model [1] were low-resolution and not very convincing. DCGAN [143] was the first approach that introduced a deconvolution layer in the generator to replace the fully connected layer, which achieved better performance in synthetic image generation. Liu et al. [144] proposed CoGAN, based on VAE, for learning joint distributions of two-domain images. This model trained a couple of GANs rather than a single one, and each was responsible for synthesizing images in one domain. The size of generated images still remained relatively small, e.g. 64 × 64 or 128 × 128 pixels.

The generation of high-resolution images was limited earlier due to memory constraints. Karras et al. [145] presented ProGAN, a training methodology for GANs, that employed an adaptive mini-batch size that progressively increased the resolution, depending on the current output resolution, by adding layers to the networks during the training process. StyleGAN [146] was an improved version of ProGAN [145]. Instead of mapping latent code z to a resolution, a Mapping Network was employed that learned to map input latent vector (Z) to an intermediate latent vector (W) which controlled different visual features. The improvement was that the intermediate latent vector was free from any distribution restriction, and this reduced the correlation between features (disentanglement). The layers of the generator network were controlled via an AdaIN operation which helped decide the features in the

Table 8 An overview of face reenactment based deepfake detection techniques

Author	Technique	Features	Best Evaluation performance	Dataset	Limitations
Handcrafted					
Matern et al. [81]	MLP, Logreg	16-D texture energy based features of eyes and teeth [82]	▪ AUC=.823 (MLP) ▪ AUC=.866 (LogReg)	FF++	▪ Only applicable to face images with open eyes and clear teeth.
Agarwal et al. [83]	SVM Classifier	16 AU's using OpenFace2 toolkit	▪ AUC=98%	Own dataset.	▪ Degraded performance in cases where a person is looking off-camera.
Amerini et al. [138]	VGG16, ResNet	Optical flow fields	Accuracy=81.61% (VGG16), 75.46% (ResNet)	FF++	▪ Very few results are reported
Deep Learning					
Sabir et al. [91]	CNN/RNN	CNN features	Accuracy=94.35%	FF++	▪ Results are reported for static images only.
Afchar et al. [92]	MesoInception-4	Deep features (DF)	TPR=81.3%	FF++	▪ Performance degrades on low quality videos.
Nguyen et al. [93]	CNN	Deep features	Accuracy=92.50%	FF++	▪ Degraded detection performance for unseen cases.
Stehouwer et al. [94]	CNN	Deep features	Accuracy=99.4%	Diverse Fake Face Dataset (DFFD)	▪ Computationally complex due to large feature vector space.
Rossle et al. [95]	SVM+CNN	Co-Occurance matrix + DF	Accuracy=86.86%	FF++	▪ Low performance on compressed videos.

output layer. Compared to [1, 143, 144], StyleGAN [146] achieved state-of-the-art high resolution in the generated images i.e., 1024 × 1024, with fine detail. StyleGAN2 [147] further improved the perceived image quality by removing unwanted artifacts, such as a change in gaze direction and teeth alignment with the facial pose. Huang et al. [148] presented a Two-Pathway Generative Adversarial Network (TP-GAN) that could simultaneously perceive global structures and local details, like humans, and synthesized a high-resolution frontal view facial image from a single ill-posed face image. Image synthesis using this approach preserved the identity under large pose variations and illumination. Zhang et al. [149] introduced a self-attention module in convolutional GANs (SAGAN) to handle global dependencies, and thus ensured that the discriminator can accurately determine the related features in distant regions of the image. This work further improved the semantic quality of the generated image. In [150], the authors proposed BigGAN architecture, which used residual networks to improve image fidelity and the variety of generated samples by increasing the batch size and varying latent distribution. In BigGAN, the latent distribution was embedded in multiple layers of the generator to influence features at different resolutions and levels of the hierarchy rather than just adding to the initial layer. Thus, the generated images were photo-realistic and very close to real-world images from the ImageNet dataset. Zhang et al. [151] proposed a stacked GAN (StackGAN) model to generate high-resolution images (e.g., 256 × 256) with details based on a given textual description. In [152], spatial and channel attention layers were added to the

Fig. 10 Improvement in the quality of synthetic faces generated by variations on GANs. In order, the images are from papers by Goodfellow et al. (2014) [1], Radford et al. (2015) [143], Liu et al. (2016) [144], Karras et al. (2017) [145], and Style-based (2018 [146], 2019 [147])

Table 9 An overview of face synthesis deepfake generation techniques

Reference	Technique	Features	Dataset	Output Quality	Limitations
Liu et al. [144]	CoGAN	Deep Features	CelebA	64×64 or 128×128	▪ Generates low-quality samples
Karras et al. [145]	ProGAN	Deep Features	CelebA	1024×1024	▪ Limited control on the generated output
Karras et al. [147]	StyleGAN	Deep Features	▪ ImageNet	1024×1024	▪ Blob-like artifacts
Huang et al. [148]	TP-GAN	Deep Features	▪ LFW	256×256	▪ Lack fine details ▪ Lack semantic consistency
Zhang et al. [149]	SAGAN	Deep Features	▪ ImageNet2012	128×128	▪ Unwanted visible artifacts
Brock et al. [150]	BigGAN	Deep Features	▪ ImageNet	512×512	▪ Class-conditional image synthesis ▪ Class leakage
Zhang et al. [151]	StackGAN	Deep Features	▪ CUB ▪ Oxford ▪ MS-COCO	256×256	▪ Lack semantic consistency

generator network to improve texture learning details for super-resolution image generation.

Detection Techniques based on handcrafted Features: A lot of literature is available on image forgery detection [153–158]. As AI-manipulated data is a new phenomenon, there are few forensic techniques that work well for deepfake detection. Recently, some researchers [73, 159] have adopted the idea of employing the traditional methods of image forgery identification to detect synthesized faces, however, these approaches are unable to identify fake facial images**.** Current research has focused on new ML-based techniques. McCloskey et al. [160] present an approach to identify fake images by employing the fact that the color information is dissimilar between the real camera and synthesized samples. The color key-points from input samples are used to train the SVM for classification. This approach [160] exhibits better fake sample detection accuracy, however, it may not perform well for blurred images. Guarnera et al. [161] proposes a method to identify fake images. Initially, the EM algorithm is used to calculate the image features. The computed key-points are used to train three types of classifiers, KNN, SVM, and LDA. The approach in [161] performs well for synthesized image identification, but may not perform well for compressed images.

Techniques based on Deep Features: DL-based work such as in [162], the authors proposed a method to detect forged images by calculating the pixel co-occurrence matrices at three color channels of the image. Then a CNN model was trained to learn important features from the co-occurrence matrices to differentiate manipulated and non-manipulated content. Yu et al. [163] presented an attribution network architecture to map an input sample to its related fingerprint image. The correlation index among each sample fingerprint and model fingerprint acts as a softmax logit for classification. This approach [163] exhibited better detection accuracy, however, it may not have performed well with post-processing operations i.e. noise, compression, and blurring, etc. Marra et al. [164] proposed a study to identify GAN-generated fake images. Particularly, [164] introduced a multi-task incremental learning detection approach to locate and classify new types of GAN-generated samples without affecting the detection accuracy of the previous ones. Two solutions related to the position of the classifier were introduced by employing the iCaRL algorithm for incremental learning [165], named as Multi-Task MultiClassifier, and Multi-Task Single Classifier. This approach [164] was robust to unseen GAN-generated samples but was unable to perform well if the information on the fake content generation method is not available. Table 10 presents a comparison of the face synthesis deepfake detection techniques mentioned above.

4.1.5 Facial attribute manipulation

Generation Face attribute editing involves altering the facial appearance of an existing sample by modifying an attribute-specific region while keeping the irrelevant regions unchanged. Face attribute editing includes removing/wearing eyeglasses, changing viewpoint, skin retouching (e.g., smoothing skin, removing scars, and minimizing wrinkles), and even some higher-level modifications, such as age and gender, etc. Increasingly, people are using commercially available AI-based face editing and mobile applications such as FaceApp [5] to automatically alter the appearance of an input image.

Recently, several GAN-based approaches have been proposed to edit facial attributes, such as the color of the skin, hairstyle, age, and gender by adding/removing glasses and facial expressions in a given face. In this manipulation, the GAN takes the original face image as input and generates the edited face image with the given attribute, as shown in Fig. 11. A summary of face attribute manipulation approaches is presented in Table 11. Perarnau et al. [166] introduce the Invertible Conditional GAN (IcGAN), which uses an encoder

Table 10 An overview of face synthesis deepfake detection techniques

Author	Technique	Features	Best Evaluation performance	Dataset	Limitations
Handcrafted					
Guarnera et al. [161]	EM + (KNN, SVM, LDA)	Deep features	▪ Accuracy=99.22 (KNN) ▪ Accuracy= 99.81(SVM) ▪ Accuracy=99.61 (LDA)	CelebA	Not robust to compressed images.
McCloskey et al. [160]	SVM	Color channels	▪ AUC=70%	MFC2018	Performance degrades with blurry samples.
Deep Learning					
Nataraj et al. [162]	CNN	Deep features + co-occurrence matrices	Accuracy=99.49% Accuracy=93.42%	▪ cycleGAN ▪ StarGAN	▪ Works with static images only. ▪ Low performance for jpeg compressed images.
Yu et al. [163]	CNN	Deep features	Accuracy=99.43%	CelebA	▪ Poor performance on post-processing operations.
Marra et al. [164]	CNN+Incremental Learning	Deep features	Accuracy=99.3%	Customized	▪ Needs source manipulation technique information

in combination with cGANs for face attribute editing. The encoder maps the input face image into a latent representation and an attributes manipulation vector, and a cGAN reconstructs a face image with new attributes, given the altered attributes vector as the condition. This suffers from information loss and alters the original face identity in the synthesized image. In [167], a Fader Network is presented, where an encoder-decoder architecture is trained in an end-to-end manner which generates an image by disentangling the salient information of the image and the attribute values directly in latent space. This approach, however, adds unexpected distortion and blurriness, and thus fails to preserve the fine details of the original in the generated image.

Prior studies [166, 167] have been focused on handling image-to-image translations between two domains. These methods required different generators to be trained independently to handle translations between each pair of image domains and thus limited their practical usage. StarGAN [36], an enhanced approach, was capable of translating images among multiple domains using a single generator. A conditional facial attribute transfer network was trained via attribute classification loss and cycle consistency loss. StarGAN achieved promising visual results in terms of attribute manipulation and expression synthesis. This approach, however, added some undesired visible artifacts in facial skin, such as an uneven color tone, in the output image. The recently proposed StarGAN-v2 [168] achieved state-of-the-art visual quality of generated images as compared to [36] by adding a random Gaussian noise vector into the generator. In AttGAN [169], an encoder-decoder architecture was proposed that considered the relationship between attributes and the latent representation. Instead of imposing an attribute independent

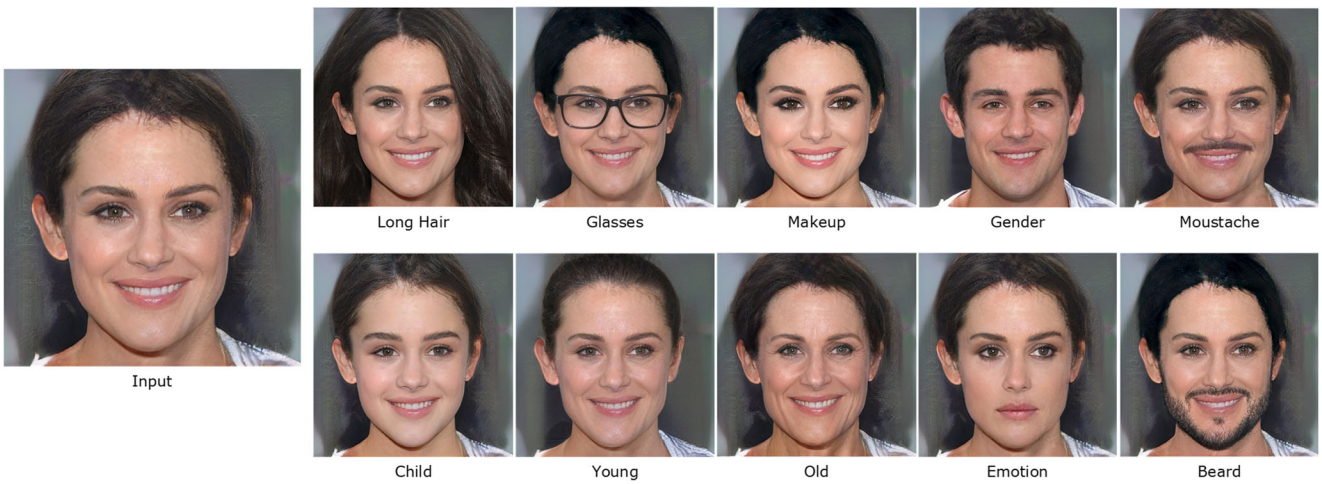

Fig. 11 Examples of different face manipulations: original sample (Input) and manipulated samples

Table 11 An overview of facial attribute manipulation-based deepfake generation techniques

Author	Technique	Features	Best Evaluation performance	Dataset	Limitations
Perarnau et al. [166]	IcGAN	▪ Deep Features	▪ CelebA ▪ MNIST	64×64	▪ Fails to preserve original face identity
Fader Network [167]	Encoder-decoder	▪ Deep Features	▪ CelebA	256×256	▪ Unwanted distortion and blurriness ▪ Fails to preserve fine details
Choi et al. [168]	StarGAN	▪ Deep Features	▪ CelebA ▪ RaFD	512×512	▪ Undesired visible artifacts in the facial skin e.g., the uneven color tone
He et al. [169]	AttGAN	▪ Deep Features	▪ CelebA ▪ LFW	384×384	▪ Generates low-quality results and adds unwanted changes, blurriness
Liu et al. [170]	STGAN	▪ Deep Features	▪ CelebA	384×384	▪ Poor performance for multiple attribute manipulation
Zhang et al. [171]	SAGAN	▪ Deep Features	▪ CelebA	256×256	▪ Lack of details in the attribute-irrelevant region
He et al. [172]	PA-GAN	▪ Deep Features	▪ CelebA	256×256	▪ undesired artifacts in case of baldness and open mouth etc.

constraint on the latent representation, like in [166, 167], an attribute classification constraint was applied to the generated image in order to guarantee the correct change of the desired attributes. AttGAN provided improved facial attribute editing results, with other facial details well preserved. However, the bottleneck layer, i.e., down-sampling in the encoder-decoder architecture, added unwanted changes and blurriness and generated low-quality edited results. Liu et al. [170] proposed the STGAN model that incorporated an attribute difference indicator and a selective transfer unit with an encoder-decoder to adaptively select and modify the encoded features. STGAN only focused on the attribute-specific region and did not guarantee good preservation of the details in attribute-irrelevant regions.

Other works introduce the attention mechanism for attribute manipulation. SAGAN [171] introduces a GAN-based attribute manipulation network to perform alteration and a global spatial attention mechanism to localize and explicitly constrain editing within a specified region. This approach preserves the irrelevant details well but at the cost of attribute correctness in the case of multiple attribute manipulation. PA-GAN [172] employs a progressive attention mechanism in a GAN to progressively blend the attribute features into the encoder features, constrained inside a proper attribute area, by employing an attention mask from high to low feature level. As the feature level gets lower (higher resolution), the attention mask gets more precise and the attribute editing becomes fine. This approach successfully performs multiple attribute manipulation, and preserves irrelevance within a single model well. However, some undesired artifacts appear in cases where significant modifications are required, such as baldness and an open mouth.

Detection Techniques based on handcrafted Features: Researchers have employed the traditional ML-based approaches for the detection of facial attributes manipulation. In [173], the author used the pixel co-occurrence matrices to compute features from the suspect samples. The extracted keypoints were used to train a CNN classifier to differentiate original and manipulated faces. The method in [173] showed better facial attribute manipulation detection accuracy, however, it may not have performed well given noisy samples. An identification approach using keypoints computed from the frequency domain, instead of employing raw sample pixels, was introduced in [174]. For each input sample, a 2D discrete fourier transformation (DFT) was applied to transform the image to the frequency domain in order to acquire one frequency sample per RGB channel. The work, [174], used an AutoGAN classifier for predicting real and fake samples. The generalization ability of the work in [174] was evaluated over unseen GAN frameworks. More specifically, they considered two GAN frameworks, namely StarGAN [36] and GauGAN [175]. The work showed better prediction accuracy for the StarGAN model, however, in the case of GauGAN, the technique faced a serious performance drop.

Techniques based on Deep Features: The research community has presented several methods to detect facial manipulations by evaluating the internal GAN pipeline. Similar work was presented in [176], where the author introduced the concept that analyzing internal neuron behavior could assist in identifying manipulated faces, as layer-by-layer neuron activation arrangements could extract a more representative set of significant image features for recognizing the original and fake faces. The proposed solution in [176], namely FakeSpotter, computed deep features by employing several DL-based face recognition frameworks, i.e., VGG-Face [177], OpenFace [178], and FaceNet [179]. The extracted features were used to train an SVM classifier to categorize fake and real faces. The solution [176] performed well for facial attributes

manipulation detection, however, it may not have performed well for samples with intense light variation.

Existing works on facial attribute manipulation have either employed entire faces or passed face patches in order to spot real and manipulated content. A face patch-based technique was presented in [180], where a Restricted Boltzmann Machine (RBM) was used to compute deep features. Then, the extracted features were used to train a two-class SVM classifier to classify real and forged faces. The method in [180] was robust to manipulated face detection, however, it was at the expense of increased computational cost. Another similar approach was proposed in [181], where a CNN-based keypoint extractor was presented. The CNN approach comprised six convolutional layers, along with two fully connected layers. Additionally, residual connections were introduced which allowed the ResNet frameworks to compute the deep features from the input samples. Finally, the calculated features were used to train an SVM classifier to predict real and manipulated faces. The approach in [181] showed better manipulation identification performance, however, it did not perform well in terms of various post-processing attacks, i.e., noise, blurring, intensity variations, and color changes. Some researchers have employed the use of entire faces rather than face patches in order to detect facial attribute manipulation in visual content. One such work was presented by Tariq et al. [182], where several DL-based frameworks, i.e., VGG-16, VGG-19, ResNet, and XceptionNet, were trained on suspect samples in order to locate facial attribute forgeries. The work in [182] showed better face attribute manipulation detection, however its performance declined in real-world scenarios. Some authors used attention mechanisms to further enhance training in the attribute manipulation detection systems. Dang et al. [183] introduced a framework to identify several types of facial manipulation. This framework employed attention mechanisms in order to enhance feature map calculation in CNN frameworks. Two different methods of attribute manipulation generation were taken into account: i) fake samples generated using the publicly available FaceApp software, with various available filters, and ii) fake samples generated with the StarGAN network. The work [183] is robust to face forgery detection, however, at the expense of high computational cost.

Wang et al. [170] proposed a framework to detect manipulated faces which encompassed two classification steps: local and global predictors. A Dilated Residual Network (DRN) model was used as a global predictor to identify real and fake samples, while optical flow fields were utilized for local predictions. The approach in [170] worked well for face attribute manipulation identification but required extensive training data. Similarly, [164] proposed a DL-based framework, XceptionNet, for the detection of face attribute forgeries. However, the method in [164] suffered from high computational cost. Rathgeb et al. [184] introduced Photo Response Non-Uniformity (PRNU). In this method, scores gathered after performing an analysis of spatial and spectral features, computed from the PRNU patterns from entire image samples, were fused. The approach [184] was able to robustly differentiate between bonafide and retouched facial samples, however accuracy was lacking.

Many of these DL-based methods achieve near-perfect accuracy, as shown in Table 12, however this accuracy appears to be largely due to the presence of GAN fingerprints in the manipulated samples. Newer research focuses on detection in samples where the GAN signatures have been removed, and this has proven to be challenging for previously high-performing frameworks. Hence, the research community needs to develop strategies that are resistant to such attacks.

4.1.6 Discussion of visual manipulation methods

Generation Deepfake generation has advanced significantly in recent years. The high quality of generated images across different visual manipulation categories (face-swap, face-reenactment, lip-sync, entire face synthesis, and attribute manipulation) has made it increasingly difficult for human eyes to differentiate between fake and genuine content. Among the significant advances are: (i) unpaired self-supervised training strategies avoid the requirement for extensive labeled training data, (ii) the addition of AdaIN layers, pix2pixHD network, self-attention modules, and feature disentanglement for improved synthesized faces, (iii) one/few-shot learning strategies enable identity theft with limited target training data, (iv) the use of temporal discriminators and optical flow estimation to improve coherence in the synthesized videos, (v) introduction of a secondary network for seamless blending of composites in order to reduce boundary artifacts, (vi) the use of multiple loss functions to handle different tasks, such as conversion, blending, occlusion, pose, illumination, etc., for improved final output, and (vii) the adoption of perceptual loss with pretrained VGG-Face network dramatically enhanced synthesize facial quality. Current deepfake systems have a few limitations, e.g., in facial reenactment generation techniques frontal poses are always used to drive and create the content. As a result, reenactment is restricted to a somewhat static performance. Currently, Face-swapping onto the body of a lookalike is performed to achieve facial reenactment, however, this approach has limited flexibility because having a good match is not always achievable with the current technology. Moreover, face reenactment depends on the driver's performance to portray the target identity personality. Recently, there has been a trend towards identity-independent deepfake generation models. Another development is real-time deepfakes that allow face swapping in video chats. Real-time deepfakes at 30fps have been achieved in works such as [67, 106]. The next generation of deepfakes are expected to utilize video

Table 12 An overview of facial attribute manipulation based deepfake detection techniques

Author	Technique	Features	Best Evaluation performance	Dataset	Limitations
Hand-crafted					
[173]	co-occurrence matrices along with CNN	Co-Occurrence matrix	Accuracy=99.4%	Private dataset.	▪ Its evaluation performance reduces over noisy images.
[174]	GAN Discriminator	Frequency domain features	Accuracy =100%	Private dataset.	▪ The technique faces serious performance degradation for GauGAN framework-based face attribute manipulations.
Deep Learning					
[176]	FakeSpoter	Deep features	Accuracy=84.7%	Private dataset	▪ Detection performance decreases when the samples have significant light variation.
[180]	RBM along with the SVM classifier	Deep features	Accuracy=96.2% Accuracy=87.1%	Private dataset Private dataset (Celebrity Retouching, ND-IIITD Retouching)	▪ This method suffers from the high computational cost.
[181]	CNN+SVM	Deep features	Accuracy =99.7%	Private dataset	▪ Results are reported for post-processing attacks.
[182]	CNNs	Deep features	AUC=74.9%	Private dataset	▪ Performance degrades in real-world scenarios.
[183]	Attention Mechanism along with CNN	Deep features	AUC=99.9%	DFFD	▪ This work is computationally complex.
[170]	DRN	Deep features	Average precision= 99.8%	Private dataset	▪ The approach should be evaluated over a standard dataset.
[164]	Incremental Learning along with the CNN	Deep features	Accuracy =99.3%	Private dataset	▪ This work is inefficient.
[184]	Score-Level Fusion	PRNU Features	EER=13.7%	Private dataset	▪ The work needs to improve the classification accuracy.

stylization techniques to generate target manipulated content with projected expression and mannerism. Although, existing deepfakes are not perfect, the rapid development of high-quality real/fake image datasets promote deepfake generation research.

Detection In this subsection, we presented a summary of the work performed for visual deepfakes detection. Based on the in-depth analysis of various detection approaches, we concluded that most of the existing detection work is based on employing a DL-based approach and shows a robust performance approaching 100%. The main reason for the accuracy of models is the presence of fingerprint information, visible artifacts in the audiovisual manipulated samples. However, more recently researchers have presented approaches which removed the information from the forged samples, which is proving to be a challenge even for high-performing attribute manipulation detection frameworks. It has been observed that most of the existing detection techniques perform well on face swap detection, and are relatively easily able to identify when the entire face is swapped with the target identity, which usually leaves artifacts. However, expression swap and lip-sync are more challenging to detect as these manipulations tamper with soft biometrics of the same person's identity. For visual deepfakes detection, it has been observed that it's relatively easy for the research community to detect image-based manipulations in comparison to video-based deepfakes. Both for audio or visual deepfakes, most of the research work has used publically available datasets instead of using their own synthesized datasets. The existing works have reported robust performance for visual deepfake detection but has faced a serious performance drop for unseen cases, indicating a lack of generalization ability, is likely related. Moreover, these approaches are unable to definitively prove the difference between real and manipulated content, so these approaches lack explainability. Several deepfake detection methods presented in previous years have proven to be nearly unusable due to implementation complexities such as variation in datasets, configuration environment, and complicated architecture. More recently, software and online platforms such as DeepFake-o-meter [185], FakeBuster [186], and Video Authenticator (not publicly available) [187] have been introduced which are able to easily detect audio-visual manipulation and give access to the general audience. However, these platforms are in their infancy and need further development to handle emerging deepfakes.

Figure 12 groups the existing work performed for visual deepfake detection. Table 13 presents a detailed description of each category. Existing approaches have either targeted spatial and temporal artifacts left during the generation or data-driven classification. The spatial artifacts include inconsistencies [78, 81, 114, 188, 193, 201–203], abnormalities in background [160, 194, 198], and GAN fingerprints [74, 163, 204, 205]. The temporal artifacts involve detecting variation in a person's behavior [83, 88, 200], physiological signals [77, 78, 85, 89], coherence [190, 199, 206], or video frame synchronization [33, 75, 91, 138, 207, 208]. Instead of focusing on a specific artifact, some approaches are data-driven, which detect manipulations by classification [58, 73, 84, 86, 87, 92–95, 119, 123, 161, 162, 164, 189, 191, 192, 209–213] or anomaly identification [121, 122, 195, 196, 214–216]. Moreover, in Fig. 12, the * references show the DL-based approaches employed for deepfake detection, while others show the hand-coded feature extraction methods.

4.2 Audio manipulations

AI-synthesized audio manipulation is a type of deepfake that can clone a person's voice and depict that voice saying something that the person never said. Recent advancements in AI-synthesized algorithms for speech synthesis and voice cloning have shown the potential to produce realistic fake voices that are nearly indistinguishable from genuine speech. These algorithms can generate synthetic speech that sounds like the target speaker, based on text or samples of the target speaker, with highly convincing results [59, 217]. Synthetic voice is widely adapted for the development of different applications, such as automated dubbing for TV and film, chatbots, AI assistants, text readers, and personalized synthetic voices for vocally handicapped people. Aside from this, synthetic/fake voices have become an increased threat to voice biometric systems [218] and used for malicious purposes, such as political gains, fake news, or fraud as well [14, 58]. More complex audio synthesis could combine the power of AI with manual editing. For example, neural network-powered voice synthesis models, such as Google's Tacotron [56], Wavenet [55], or AdobeVoco [219], can generate realistic and convincing fake voices that resemble the victim's voice. Later on, audio editing software, e.g. Audacity [6], can be used to integrate the original and synthesized audio to make more convincing fakes.

AI-based impersonation is not limited to visual content; recent advancements in AI-synthesized fake voices are assisting the creation of highly realistic deepfakes video [37]. These developments in speech synthesis have shown a potential to produce realistic and highly natural sounding audio deepfakes, exhibiting a real threat to society [14]. Combining synthetic audio content with visual manipulation can make deepfake videos significantly more convincing and increase their impact [37]. Despite much progress, synthesized speech still lacks some aspects of voice quality, like

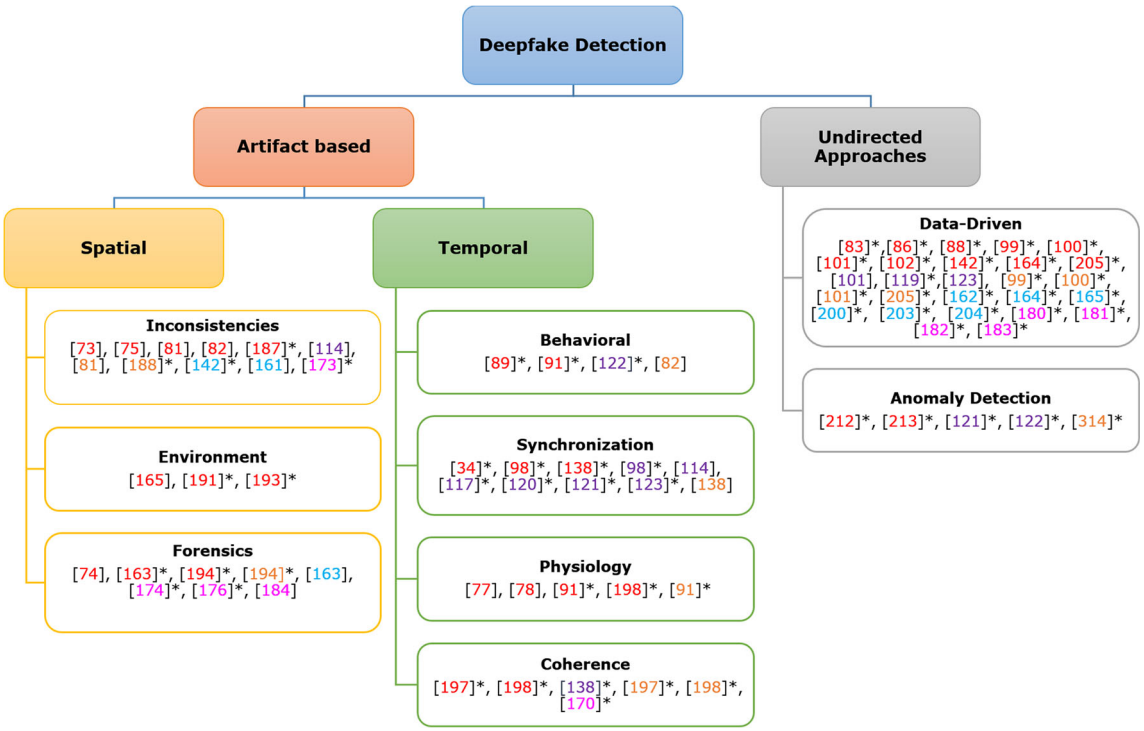

Fig. 12 Categorization of visual deepfake detection techniques (The red color shows Face-Swap detection approaches, purple for Face-Reenactment, Orange for lip-syncing, Blue for facial image synthesis, and pink for facial attribute manipulation detection techniques, where * shows deep-learning based approaches)

Table 13 Description of classification categories for existing deepfake detection methods

Inconsistencies	Visible artifacts within the frame such as inconsistent head poses and landmarks etc.
Environment	Abnormalities in the background such as lighting and other details.
Forensics	GAN fingerprints left during the generation process.
Behavioral	Monitoring abnormal gestures and facial expressions.
Synchronization	Temporal consistency such as inconsistencies between adjacent frames/modality.
Physiology	Lack of biological signals such as eye blinking patterns and heart rate
Coherence	Missing optical flow field and artifacts such as flickering and jitter between frames
Classification	End-to-end CNN based data-driven models
Anomaly Detection	Outliers identification such as reconstructing real images and comparing to the encoded image. They are used to see unknown creation methods.

expressiveness, roughness, breathiness, stress, and emotion, specific to a target identity [220]. The AI research community is making a concerted effort to overcome these challenges and produce human-like voice quality with high speaker similarity.

Two distinct modalities for audio deepfakes are text-to-speech (TTS) synthesis and voice conversion (VC). TTS synthesis is a technology that can synthesize a natural-sounding sample of any speaker based on the given input text [221]. VC is a technique that modifies the audio waveform of a source speaker to a sound similar to the target speaker's voice [222]. A VC system takes the recording of an individual as a source and creates a deepfake audio in the target's voice. It preserves the linguistic and phonetic characteristics of the source sample and changes them to that of the target speaker. TTS synthesis and VC represent a genuine threat when used maliciously as both generate completely synthetic computer-generated voices that are nearly indistinguishable from genuine speech. Moreover, cloned replay attacks [13] impose a potential risk for voice biometric devices because the latest speech synthesis techniques can produce a vocal sample with high speaker similarity [223]. This section lists the latest progress in speech synthesis including TTS and VC techniques as well as detection strategies.

4.2.1 TTS voice synthesis

TTS is a decades-old technology which can synthesize a natural-sounding voice from a given input text, and thus enables a voice to be used for better human-computer interaction. The initial research on TTS synthesis technology was done using the methods of speech concatenation or parameter estimation. The concatenative TTS systems are based on separating high-quality recorded speech into small fragments followed by concatenation into a new speech. In recent years, this method has become outdated and unpopular as it is not scalable or consistent. In contrast, parametric models map text to the salient speech parameters and convert them into an audio signal using vocoders. Later on, the deployment of deep neural networks has gradually become a dominant method for speech synthesis that achieves much better voice quality. These methods include the Neural vocoders [55, 221, 224], GANs [225–227], autoencoders [228], autoregressive models [229–231], and other emerging techniques [228, 232–236] which have promoted the rapid development of the speech synthesis industry. Figure 13 shows the principle design of modern TTS methods.

The significant developments in voice/speech synthesis are WaveNet [55], Tacotron [56], and DeepVoice3 [224], which can generate realistic sounding synthetic speech from an input text to provide an enhanced interaction experience between humans and machines. Table 14 presents an overview of the state-of-the-art speech synthesis methods. WaveNet [55], developed by DeepMind in 2016 utilizes raw audio waveforms by processing acoustic features, i.e., spectrograms, through a generative framework that is trained on actual recorded speech. Parallel WaveNet has been introduced to enhance sampling efficacy and produce high-fidelity audio signals [231]. Another DL based using a variant of WaveNet, Deep

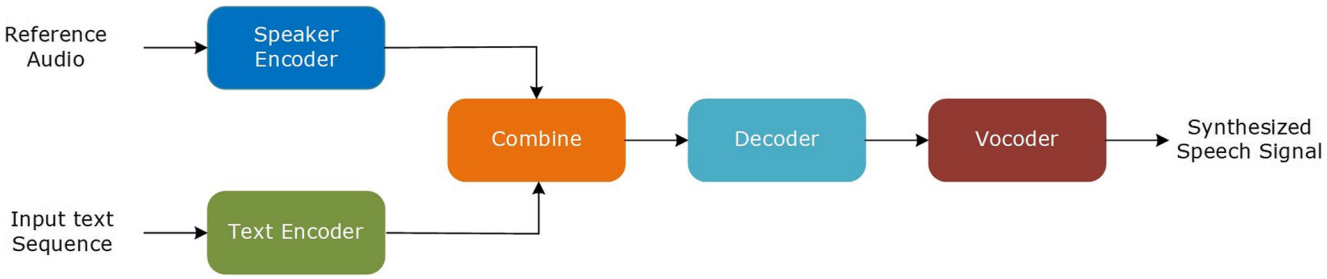

Fig. 13 Workflow diagram of the latest TTS systems

Table 14 An overview of the state-of-the-art text-based speech synthesis techniques

Methods	Technique	Features	Dataset	Limitations
WaveNet [55]	Deep neural network	▪ linguistic features ▪ fundamental frequency (log F0)	▪ VCTK (44 hrs.)	▪ Computationally complex
Tacotron [56]	Encoder-Decoder with RNN	▪ Deep features	Private (24.6 hrs.)	▪ Costly to train the model
Deep Voice 1[57]	Deep neural networks	▪ linguistic features	Private (20 hrs.)	▪ Independent training of each module leads to a cumulative error in synthesized speech
Deep Voice 2 [237]	RNN	▪ Deep features	VCTK (44 hrs.)	▪ Costly to train the model
DeepVoice3 [224]	Encoder-decoder	▪ Deep features	▪ Private (20 hrs.) ▪ VCTK (44 hrs.) ▪ LibriSpeech ASR (820 hrs.)	▪ Does not generalize well for unseen samples.
Parallel WaveNet [231]	Feed-forward neural network with dilated causal convolutions	▪ linguistic features	Private	▪ Requires a large amount of the target's speech training data.
VoiceLoop [230]	Fully-connected neural network	▪ 63-dimensional audio features	▪ VCTK (44 hrs.) ▪ Private	▪ Low ecological validity
Tacotron2[238]	▪ Encoder-decoder	▪ linguistic features	▪ Japanese speech corpus from the ATR Ximera dataset (46.9 hrs.)	▪ Lack of real-time speech synthesis
Arik et al. [59]	Encoder- decoder	▪ Mel spectrograms	▪ LibriSpeech (820 hrs.) ▪ VCTK (44 hrs.)	▪ Low performance for multi-speaker speech generation in the case of low-quality audio
Jia et al. [233]	Encoder-decoder	▪ Mel spectrograms	▪ LibriSpeech (436 hrs.) ▪ VCTK (44 hrs.)	▪ Fails to attain human-level naturalness ▪ Lacks in transferring the target accent, prosody to synthesized speech
Luong et al. [228]	Encoder-decoder	▪ Mel spectrograms	▪ LibriSpeech (245 hrs.) ▪ VCTK (44 hrs.)	▪ Low performance in the case of noisy audio samples
Chen et al. [235]	Encoder + deep neural network	▪ Mel spectrograms	▪ LibriSpeech (820 hrs.) ▪ private	▪ Low performance in the case of a low-quality audio sample
Cong et al. [236]	Encoder-decoder	▪ Mel spectrograms	▪ MULTI-SPK ▪ CHiME-4	▪ Lacks in synthesizing utterances of a target speaker

Voice 1 [57], puts each module containing an audio signal, voice generator, or a text analysis front-end through a related NN model. Due to the independent training of each module, however, it is not a real end-to-end speech synthesis system.

In 2017, Google introduced tacotron [56] an end-to-end speech synthesis model. Tacotron could synthesize speech from given <text, audio> pairs and thus generalized well to other datasets. Similar to WaveNet, the Tacotron framework was a generative framework comprised of a seq2seq model that contained an encoder, an attention-based decoder, and a post-processing network. Even though the Tacotron model attained better performance it had one potential limitation i.e., it must employ multiple recurrent components. The inclusion of these units made it computationally inefficient so it required high-performance systems for model training. Deep Voice 2 [237] combined the capabilities of both the Tacotron and WaveNet models for voice synthesis. Initially, Tacotron was employed for converting the input text to a linear scale spectrogram, later converted to voice through the WaveNet model.

In [238], Tacotron2 is introduced for vocal synthesis and it exhibits an impressively high mean opinion score, very similar to human speech. Tacotron2 consists of a recurrent sequence-to-sequence keypoint estimation framework that maps character embedding to mel-scale spectrograms. To deal with the time complexities of recurrent unit-based speech synthesis models, a new, fully-convolutional character-to-spectrogram model named DeepVoice3 is presented in [224]. The Deep Voice 3 model is faster than its peers due to performing fully parallel computations. Deep Voice 3 is comprised of three main modules: i) an encoder that accepts text as input and transforms it into an internal learned form, ii) a decoder that converts the learned representations in an autoregressive manner, and iii) a post-processing, fully convolutional network that predicts the final vocoder parameters.

Another model for voice synthesis is VoiceLoop [230], which uses a memory framework to generate speech from voices unseen during training. VoiceLoop builds a phonological store by executing a shifting buffer as a matrix. Text strings are characterized as a list of phonemes that are later decoded in short vectors. The new context vector is produced by assessing the encoding of the resulting phonemes and summing them together. The above-mentioned powerful end-to-end speech synthesizer models [224, 238] have enabled the production of large-scale commercial products, such as Google Cloud TTS, Amazon AWS Polly, and Baidu TTS. All these projects aim to attain a high similarity between synthesized and human voices.

The latest TTS systems can convert given text to human speech with a particular voice identity. Using generative models, researchers have built voice imitating TTS models that can clone the voice of a particular speaker in real-time using a few samples of reference speech [233, 234]. The key distinction between voice cloning and speech synthesis systems is that the former focuses on preserving the characteristics of the specific identity speech attributes while the latter lacks this feature to maintain the quality of the generated speech [228]. Various AI-enabled voice cloning online platforms are available, such as Overdub,[1] VoiceApp[2], and iSpeech,[3] which can produce synthesized voices that closely resemble the target's speech, and give the public access to this technology. Jia et al. [233] proposes a Tacotron 2 based TTS system capable of producing multi-speaker speech, including those unseen during training. The framework consists of three independently trained neural networks. The findings show that although the synthetic speech resembles a target speaker's voice it does not fully isolate the voice of the speaker from the prosody of the audio reference. Arik et al. [59] propose a Deep Voice 3-based technique comprised of two modules: speaker adaptation and speaker encoding. For speaker adaptation, a multi-speaker generative framework is fine-tuned. For speaker encoding, an independent model is trained to directly infer a new speaker embedding, which is applied to the multi-speaker generative model.

Loung et al. [228] propose a speech generation framework that can synthesize a target-specific voice, either from input text or a reference raw audio waveform from a source speaker. The framework consists of a separate encoder and decoder for text and speech, and a neural vocoder. The model is jointly trained with linguistic latent features, and the speech generation model learns a speaker-disentangled representation. The obtained results achieve quality and speaker similarity to the target speaker; however, it takes almost 5 minutes to produce the cloned speech. Chen et al. [235] propose a meta-learning approach using the waveNet model for voice adaption with limited data. Initially, speaker adaptation is computed by fine-tuning the speaker embedding. Then, a text-independent parametric approach is applied whereby an auxiliary encoder network is trained to predict the embedding vector of new speaker. This approach performs well on clean and high-quality training data however the presence of noise deviates the speaker encoding and directly affects the performance of the synthesized speech. In [236], the authors propose a seq2seq multi-speaker framework with domain adversarial training to produce a target speaker voice from only a few available noisy samples. The results show improved naturalness in the synthetic speech. However, similarity still remains challenging to achieve due to an inability to transfer target accents and prosody to synthesized speech with a limited amount of low-quality speech data.

Different GAN-based architectures have been applied to process and generate high-quality speech in audio synthesis. Notable works include WaveGAN [239], GAN-TTS [225], MelGAN [226], and Hifi-GAN [227]. Some works introduce GAN-based vocoders that focus on producing high-quality speech while maintaining controllability. In [225], the authors introduce GAN-TTS, a linguistic to waveform generation model using a GAN. It is based on a conditional feed-forward generator network that generates a raw speech waveform, and an ensemble of discriminator networks that use multi-frequency random windows to assess synthesized speech. In [226], the authors introduce Mel-GAN, a dilated convolutional structure to enlarge the receptive field in order to better simulate long-range correlation in the waveform sequences. A multi-scale discriminator network is used with a feature matching loss over the feature map of real and synthetic audio. In [227], a generator is based on a multi-receptive field fusion module that processes many patterns of varying durations simultaneously. Multiple sub-discriminators are used to individually evaluate different periodic portions of the input waveform. The loss function, similar to [226], is used to compute the distance between the produced waveform's mel-spectrogram and ground truth. The HiFi-GAN can efficiently synthesize speech that closely resembles natural speech, however, for high-quality speech synthesis, it requires model fine-tuning and respective ground-truth data.

Aside from naturalness, expressiveness is an important factor that differentiates synthesized speech from human speech. Numerous factors influence the expressiveness of a synthetic voice, including content, timbre, phonation, style, emotion, and others. An expressive TTS requires a one-to-many mapping that matches voice variants to a text selection in terms of pitch, loudness, time, and speaker accent. In [240], a feed-forward transformer network that generates mel-spectrograms from text and then synthesizes speech is proposed. Because a mel-spectrogram sequence is substantially

[1] https://www.descript.com/overdub
[2] https://apps.apple.com/us/app/voiceapp/id1122985291
[3] https://www.ispeech.org/apps

lengthier than its corresponding phoneme sequence, a monotonic alignment search is employed to extract a duration that aligns both text and speech and provides better control over the vocal speed and prosody. Similarly, work in [229] employs a fully convolutional network to generate mel-spectrograms for speech synthesis, along with a positional attention mechanism that aligns speech and text sequences. Kim et al. [232] introduce Glow-TTS, a Flow-based model for the generation of mel-spectrograms. This model uses a self-attention mechanism to internally learn mappings between the text and the latent representation of speech by using properties of flow and dynamic programming. The Glow-TTS model synthesizes natural-sounding speech and provides better control over the synthesized speech, such as speaking rate or pitch but it involves a huge number of training parameters. In addition, computing average mel-spectrograms from input leads to low-quality and less expressive synthesized speech because it lacks the ability to capture the expression details of every single utterance. Therefore, more efficient approaches that can better model different variations of speech are required to improve the expressiveness of the synthesized speech.

4.2.2 Voice conversion

Voice Conversion (VC) is a speech-to-speech synthesis technology that manipulates an input voice to sound like the target voice identity while maintaining the linguistic content of the source speech. VC has numerous applications in real life, including expressive voice synthesis, personalized speech speaking assistants, adaptive equipment for vocally impaired people, voice dubbing for the entertainment industry, and many others [222]. The recent development of anti-spoofing for automated speaker verification [218] included VC systems for the generation of spoofing data [241, 242].

In general, to perform VC high-level features of the speech, e.g., voice timbre and prosody characteristics are used. Voice timber is concerned with spectral properties of the vocal tract during phonation, whereas prosody relates to suprasegmental characteristics, i.e., pitch, amplitude, stress, and duration. Multiple Voice Conversion Challenges (VCC) have been held to encourage the development of VC generation techniques and improve the quality of converted speech [137, 241, 242]. Earlier VCC aimed to convert source speech to target speech by using non-parallel and parallel data [137, 241] but more recent [242] focused on the development of cross-lingual VC techniques, where the source speech is converted to sound like target speech using nonparallel training data across different languages.

In earlier studies VC techniques were based on spectrum mapping using paired training data, where speech samples from both the source and target speaker speaking the same linguistic content are required for conversion. Methods using GMMs [243, 244], partial least square regression [245], exemplar-based [246] techniques and others [247–249] were proposed for parallel spectral modeling. These [243–246] were "shallow" VC methods that transformed source speech spectral features directly in the original feature space. Nakashika et al. [247] proposed a speaker-dependent sequence modeling method based on and RNN to capture temporal correlations in an acoustic sequence. In [248, 249], a deep bidirectional LSTM (DBLSTM) was employed to capture long-range contextual information and generate high-quality converted speech. DNN based methods [247–249] efficiently learned feature representation for feature mapping in parallel VC, however they require large-scale paired source and target speaker utterance data for parallel training that is not feasible for practical applications in the real world.

VC methods for non-parallel (unpaired) training data are proposed to achieve VC for multiple speakers with different languages. Powerful VC techniques based on neural networks [250], vocoders [251, 252], GANs [253–259], and VAE [260–262] are introduced for non-parallel spectral modeling. Auto-encoder-based approaches attempt to learn disentangled speaker information from linguistic content and independently convert the speaker's identity. The work in [262] investigates the quality of a learned representation by comparing different auto-encoding methods. It shows that a combination of a Vector Quantized VAE and a WaveNet [55] decoder better preserves speaker invariant linguistic content and retrieves information discarded by the encoder. However, VAE/GAN-based methods tend to over smooth the transformed features because of dimensionality reduction bottleneck. Thus, the low-level information such as pitch contour, noise, and channel data is lost, that results in buzzy-sounding converted voices.

Recent GAN-based approaches, such as CycleGAN [253–256], VAW-GAN [257], and StarGAN [258], attempt to achieve high-quality transformed speech using non-parallel training data. Studies [254, 258] demonstrate state-of-the-art performance for multilingual VC in terms of both naturalness and similarity, however, performance is speaker-dependent and degrades for unseen speakers. Neural vocoders have rapidly become the most popular vocoding approach for speech synthesis due to their ability to generate human-like speech [224]. A vocoder learns to generate audio waveform from acoustic features. The study [252] analyzes the performance of different vocoders and shows that parallel-WaveGANs [239] can effectively simulate the data distribution of human speech, with acoustic characteristics, for VC. The performance, however, is still restricted for unseen speaker identity and noisy samples [217]. Recent VC methods based on TTS, like AttS2S-VC [263], Cotatron [264], and VTN [265] use text labels to synthesize speech directly by extracting aligned linguistic characteristics from the input voice. This ensures that the converted speaker and the target speaker's identity

are the same. However, these methods necessitate the use of text labels, which are not always readily accessible.

Recently, one-shot VC techniques [266, 267] are presented. In contrast to earlier techniques, the data samples of source and target speakers are not required to be seen during training. Furthermore, just one utterance from the source and target speakers is required for conversion. The speaker embedding is extracted from the target speech, which can control the speaker identity of the converted speech independently. Despite these advancements, the performance of few-shot VC techniques for unseen speakers is not stable [268]. This is primarily due to the inadequacy of speaker embedding extracted from a single speech sample from an unseen speaker [269] which significantly impacts the reliability of one-shot conversions. Other work [270–272] adopts zero-shot VC, where the source and target speakers are unseen during training, and also without re-training the model by employing an encoder-decoder architecture. The encoder extracts style and content information into style embedding and content embedding, then the decoder constructs a speech sample by combining style and content embedding. The zero-shot VC scenario is attractive because no adaptative data or parameters are required, however the adaptability quality is insufficient, especially when the target and source speakers are unseen, diverse, or noisy [268]. The summary of voice conversion techniques discussed above are presented in Table 15.

4.2.3 Audio deepfake detection

Due to recent advances in TTS [55, 224] and VC [268] techniques, audio deepfakes have become an greater threat to voice biometric interfaces and society [58]. In the field of audio forensics, there are several approaches for identifying spoofed audio. Existing works, however, fail to fully tackle the detection of synthetic speech [276]. In this section, we review the approaches proposed for the detection of audio deepfakes. Table 16 presents the comparison of audio deepfake detection techniques using both handcrafted and deep features.

Techniques based on handcrafted Features: Yi et al. [278] presented an approach to identify TTS-based manipulated audio content. In [278] hand-crafted features Constant Q cepstral coefficients (CQCC) were used to train GMM and LCNN classifiers to detect TTS synthesized speech. This approach exhibits better detection performance for fully synthesized audio, however performance degrades rapidly for partially synthesized audio clips. Li et al. [277] propose a modified ResNet model Res2Net. They evaluate the model using different acoustic features and obtain the best performance using CQT features. This model exhibits better audio manipulation detection performance, however its generalization ability needs further improvement. In [283], mel-spectrogram features with ResNet-34 are employed to detect spoofed speech. This approach works well, but its performance needs improvement. Monteiro et al. [284] propose an ensemble-based model for the detection of synthetic speech. Deep learning models, LCNNs and ResNets are used to compute deep features, which are later fused to differentiate between real and spoofed speech. This model is robust to fake speech detection, however, it needs to be evaluated on some standard datasets. Gao et al. [282] propose a synthetic speech detection approach based on inconsistencies. They employ a global 2D-DCT feature to train a residual network to detect manipulated speech. This model has better generalization ability, however, the performance degrades on noisy samples. Zhang et al. [287] propose a model to detect fake speech by using a ResNet model with a transformer encoder (TEResNet). Initially, a transformer encoder is employed to compute a contextual representation of the acoustic keypoints by considering the correlation between audio signal frames. The computed keypoints are then used to train a residual network to differentiate between real and manipulated speech. This work shows better fake audio detection performance, however, it requires extensive training data. Das et al. [279] propose a method to detect manipulated speech. Initially, a signal companding technique for data augmentation is used to increase the diversity of the training data. Then, CQT features are computed from the obtained data, which are later used to train the LCNN classifier. The method improves the fake audio detection accuracy but requires extensive training data.

Aljasem et al. [13] propose a hand-crafted, feature-based approach to detect cloned speech. Initially, sign-modified acoustic local ternary pattern features are extracted from input samples. Then, the computed keypoints are used to train an asymmetric, bagging-based classifier to categorize the samples into bona fide and fake. This work is robust to noisy cloned voice replay attacks, however, its performance needs further improvement. Ma et al. [280] present a a continual learning-based technique to enhance the generalization ability of a manipulated speech detection system. A knowledge distillation loss function is introduced in the framework to enhance the learning ability of the model. This approach is computationally efficient and can detect unseen spoofing manipulations, however, the performance has not been evaluated on noisy samples. Borrelli et al. [293] employ bicoherence features together with long-term short-term features. The extracted features are used to train three different types of classifiers: a random forest, a linear SVM, and a radial basis function (RBF) SVM. This method obtains the best accuracy with the SVM classifier. Due to handcrafted features, however, this work is not generalized to unseen manipulations. In [202] bispectral analysis is performed in order to identify specific and unusual spectral correlations present in GAN generated speech samples. Similarly, in [281] bispectral and Mel-cepstral analysis are performed in order to detect missing

Table 15 An overview of the state-of-the-art voice conversion techniques

Methods	Technique	Features	Dataset	Limitations
Ming et al. [248]	DBLSTM	F0 and energy contour	▪ CMU-ARCTIC [273]	▪ Requires parallel training data
Nakashika et al. [247]	Recurrent temporal restricted Boltzmann machines (RTRBMs)	MCC, F0, and aperiodicity	▪ ATR Japanese speech database [274]	▪ Lacks temporal dependencies of speech sequences
Sun et al. [249]	DBLSTM-RNN	MCC, F0 and Aperiodicity	▪ CMU-ARCTIC [273]	▪ Requires parallel training data
Wu et al. [250]	DBLSTM- i-vectors	19D-MCCs, Delta and Delta-Delta, F0, 400-D i-vector	▪ VCTK corpus	▪ Computationally complex
Liu et al. [251]	WaveNet vocoder	MCC and F0	▪ VCC 2018	▪ Performance degrades on inter-gender conversions
Kaneko et al. [255]	Encoder-decoder with GAN	34D-MCC, F0, and aperiodicity	▪ VCC 2018	▪ Computationally complex ▪ Domain-specific voice
Kameoka et al. [258]	Encoder-decoder with GAN	36D-MCC, F0, and aperiodicity	▪ VCC 2018	▪ Performance degrades on cross-gender conversion ▪ Low performance for unseen speakers
Zhang et al. [259]	VAW-GAN	STRAIGHT spectra [275], F0 and aperiodicity	▪ VCC2016	▪ Lacks target speaker similarity
Huang et al. [260]	Encoder-decoder	STRAIGHT spectra [275] MCCs	▪ VCC 2018	▪ Lacks multi-target VC ▪ Introduces abnormal fluctuations in generated speech
Chorowski et al. [262]	VQ-VAE, WaveNet decoder	13D-MFCC	▪ LibriSpeech ▪ ZeroSpeech 2017	▪ Over smooth and low naturalness in generated speech ▪ Increased training complexity
Tanaka et al. [263]	BiLSTM encoder-LSTM decoder	Acoustic features	▪ CMU Arctic database	▪ Requires extensive training data
Park et al. et al. [264]	Encoder-decoder	Mel-spectrogram	▪ LibriTTS ▪ VCTK dataset	▪ Requires transcribed data ▪ Lacks target speaker similarity
Huang et al. [265]	VAE-vocoder	MCCs, log F0, and aperiodicity	▪ CMU ARCTIC ▪ VCTK corpus	▪ Requires parallel training data
Lu et al. [266]	Attention mechanism in encoder-decoder	13D-MFCCs, PPGs and log F0	▪ VCTK corpus	▪ Low target similarity and naturalness in generated speech
Liu et al. [267]	Encoder and DBLSTM	19 MFCCs, log F0 and PPG	▪ VCTK corpus	▪ Low target similarity and naturalness in generated speech
Chou et al. [270]	Attention mechanism in encoder-decoder	19 MFCCs, log F0 and PPG	▪ VCTK Corpus	▪ Low quality of converted voices in case of noisy samples
Qian et al. [271]	Encoder-decoder	speech spectrogram	▪ VCTK corpus	▪ Prosody flipping between the source and the target. ▪ Not well-generalized to unseen data

durable power components in synthesized speech. The computed features are then used to train several ML-based classifiers and attained the best performance using a Quadratic SVM. These approaches [202, 281] are robust to TTS synthesized audio, however, they may not be able to detect high-quality synthesized speech. Chen et al. [285] propose a DL-based framework for audio deepfake detection. The 60-dimensional linear filter banks (LFB) are extracted from speech samples and are later used to train a modified ResNet model. This work improves fake audio detection performance but suffers from high computational cost. Huang et al. [286] present an approach for audio spoofing detection where initially, short-term zero-crossing rate and energy are utilized to identify the periods of silence in each speech signal. In the next step, the linear filter bank (LFBank) key-points are computed from the nominated segments in the relatively high-frequency domain. Lastly, an attention-enhanced DenseNet-BiLSTM framework is built to locate places wher the audio is manipulated. This method [286] avoids over-fitting at the expense of high computational cost. Wu et al. [210] introduce a novel, key-point genuinization based light convolutional neural network (LCNN) framework for the identification of manipulared speech. The attributes of the original speech are utilized to train a model using a CNN. The output is then

Table 16 An overview of audio deepfake detection techniques

Author	Technique	Features	Best Evaluation performance	Dataset	Limitations
Hand-crafted features					
Li et al. [277]	Res2Net	CQT	EER=2.502	ASVspoof2019	▪ Needs generalization improvement
Yi et al. [278]	GMM/LCNN	CQCC	EER=19.22 (GMM) EER=6.99 (LCNN)	Propriety	▪ Performance degrades for partial synthesized audio clip
Das et al. [279]	LCNN	CQT	EER=3.13	ASVspoof2019	▪ Requires extensive training data
Aljasem et al. [13]	Asymmetric bagging	Combination of MFCC, GTCC, ALTP, and spectral features	EER=5.22	ASVspoof2019	▪ Performance needs further improvement
Ma et al. [280]	CNN	60-D LFCC	EER=9.25	ASVspoof2019	▪ Performance degrades on noisy samples
AlBadawy et al. [202]	logistic regression classifier	Bispectral features	AUC=0.99	Propriety	▪ Performance may degrade on high-quality speech samples
Singh et al. [281]	Quadratic SVM	Bispectral and mel-cepstral features	Acc=96.1%	Propriety	▪ Needs evaluation on a large scale dataset
Gao et al. [282]	ResNet	2D-DCT features	EER=4.03	ASVspoof2019	▪ Performance degrades on noisy samples
Aravind et al. [283]	ResNet34	Mel-spectrogram features	EER=5.87	ASVspoof2019	▪ Performance needs improvement
Monteiro et al. [284]	LCNN/ResNet	Spectral features	EER=6.38	Propriety	▪ Results should be evaluated on a standard dataset
Chen et al. [285]	ResNet	60-dimensional LFB	EER=1.81	ASVspoof2019	▪ Computationally complex approach
Huang et al. [286]	DenseNet-BiLSTM	LFBank	EER=0.53	ASVspoof 2019	▪ Computationally complex approach.
Wu et al. [210]	LCNN	Genuine speech features	EER=4.07	ASVspoof 2019	▪ Can't deal with cloned replay attack detection.
Zhang et al. [287]	TEResNet	Spectrum features	EER=5.89 EER=3.99	ASVspoof2019 Fake-or-Real dataset [288]	▪ Requires extensive training data
Deep Learning features					
Zhang et al. [289]	ResNet-18+ OC-softmax	Deep features	EER=2.19	ASVspoof2019	▪ Performance degrades on VC.
Gomez-Alanis et al. [290]	LCG- RNN	Deep features	EER=6.28	ASVspoof 2019	▪ Fails to generalize for unseen attacks
Hua et al. [291]	Res-TSSDNet	Deep features	EER=1.64	ASVspoof2019	▪ Computationally complex
Jiang et al. [292]	CNN	Deep features	EER=5.31	ASVspoof2019	▪ Performance needs further improvement
Wang et al. [58]	DNN	Deep features	EER=0.021	Fake-or-Real dataset [288]	▪ Requires evaluation on challenging dataset

converted to an original key-point distribution closer to that of genuine speech. The transformed key-points are used with an LCNN to identify genuine and altered speech. This approach [210] is robust to synthetic speech manipulation detection. It is, however, unable to deal with cloned-replay attack detection.

Techniques based on Deep Features: Zhang et al. [289] propose a DL-based approach using ResNet-18 and a one-class (OC) softmax. They train the model to learn a feature space in which real speech can be discriminated from manipulated samples by a certain margin. This method improves the performance generalization ability against unseen attacks, however, performance degrades on VC attacks generated using waveform filtering. In [290], the authors propose a Light Convolutional Gated RNN (LCGRNN) model to compute the deep features and classify the real and fake speech. This model is computationally efficient; however, it is not generalized well to real-world examples. Hua et al. [291]

propose an end-to-end synthetic speech detection model, ResTSSDNet, for the computation of deep features and classification. This model is generalized well to unseen samples; however, this is at the expense of increased computational cost. Wang et al. [58] propose a DNN based approach with a layer-wise neuron activation mechanism to differentiate between real and synthetic speech. This approach performs well for fake audio detection, however the framework requires evaluation on challenging datasets. Jiang et al. [292] propose a self-supervised learning-based approach comprising eight convolutional layers to compute deep features and classify original and fake speech. This work is computationally efficient but detection accuracy needs enhancement. Malik et al. [294] propose a CNN for cloned speech detection. Initially, audio samples are converted to spectrograms on which a CNN framework is used to compute deep features and classify real and fake speech samples. This approach shows better fake audio detection accuracy but performance degrades on noisy samples. Similarly, in [295], a spatial-temporal CNN model is proposed to process mel-spectrogram sequences in order to identify given audio sample as real or fake.

Most of the above-mentioned fake speech detection have been evaluated on the ASVspoof2019 [218] dataset, however, the recently launched ASVspoof2021 [296] has opened new challenges for the research community. This dataset introduces a separate speech deepfake category that includes highly compressed TTS and VC samples without speaker verification.

4.2.4 Discussion on audio manipulation methods

Generation Extensive work has been presented on the generation of correct and natural speech for real-world applications, however, several areas require further improvement. A good speech synthesis model should produce a both realistic and clear voice. For this reason, existing works have tried to improve the articulation and genuineness of speech synthesis [55–57]. In recent years, the quality of synthetic voice has improved significantly via the use of deep learning techniques. The significant improvements include voice adaptation [59, 235], one/few-shot learning [266, 267], self-attention network [270], and cross-lingual voice transfer [254, 258]. However, the ability to produce a more human-like natural-sounding speech in the presence of noise remains challenging. Another main aim of speech synthesis techniques is to deploy a lightweight model that requires less training data [231]. Some of the work on this subject is presented in [270–272], however, these approaches lack the ability to maintain naturalism in synthesized speech. Therefore, there is a need to develop an efficient and effective speech synthesis model that requires less training data and resources which is also able to maintain realism. Furthermore, an audio signal is generated with a sampling frequency less than 16 kHz, it causes a considerable drop in the perceived speech quality [297]. The quality of synthesized speech can be improved by increasing the sampling rate. Some of the existing works suffer from word repetition, skipping, long pause or babbling problems, which cause a loss in the intelligibility of the generated speech [229–231]. To address this problem, existing models have introduced style/prosody transfer to generate more expressive voices [229, 232, 240]. Moreover, speech synthesis techniques to maintain the audio of the specific target are further required to be explored [235, 236]. Therefore, there is a need to develop such systems that can efficiently adapt to a specific target with limited data and high efficiency.

Detection We have presented a detailed literature review of the techniques employed for the detection of synthesized speech in Section 4.2.3. Most of the existing detection approaches are based on the employment the hand-coded features for the detection of altered speech [277–285, 287, 293]. Some additional works have utilized end-to-end training models to detect audio manipulation [58, 292], while others have employed both hand-coded and deep features in a training module for speech synthesis detection [286]. Only a few techniques are focused on the detection of more than one type of audio deepfake, e.g., TTS and VC [58, 281]. In the realm of audio manipulation, VC detection has proven more challenging compared to TTS [218]. Several works have used CNN-based methods [292, 295], ensemble methods based on different feature representations [284], or methods that detect unusual aspects in human speech [202, 281]. Several variants of the ResNet model have used deep features to detect audio spoofing [289, 291]. However, one of the limitations of the existing works is the lack of generalization of the detection models. The performance significantly degrades when evaluated on unseen or samples generated with different manipulation methods [202, 290]. Lastly, an additional limitation of the existing techniques is detection performance with limited training data and computational resources [289–291].

5 Deepfake datasets

To analyze the detection accuracy of proposed methods it is of utmost importance to have a good and representative dataset for performance evaluation. Moreover, the techniques should be validated across datasets to show their ability to generalize. Therefore, researchers have put in significant effort over the years preparing standardized datasets for manipulated video and audio content. In this section, we present a detailed review of the standard datasets that are currently used to evaluate the performance of audio and video deepfake detection techniques. Tables 17 and 18 show a comparison of available video and audio deepfake datasets respectively.

Table 17 Comparison of video deepfakes detection datasets

	UADFV [74]	DF-TIMIT [191]	FF++ [95]	Celeb-DF [194]	DFDC-preview [298]	DF [299]	WDF [300]	FN [301]	FakeAV-Celeb [302]
Released	Nov, 2018	Dec, 2018	Jan, 2019	Nov, 2019	Oct, 2019	June, 2020	Oct, 2020	March, 2021	Aug, 2021
Total videos	98	620	4000	1203	5250	60,000	7314	221,247	20,000
Real content	48	–	1000	408	1131	10,000	3805	–	500
Fake content	48	620	3000	795	4119	50,000	3509	–	19,500
Tool/ technology used for fake content generation	Fakeapp application [42]	Faceswap- GAN [65]	Deepfake, CG-manipulations	Deepfake	Unknown	DF-VAE	Collected from internet	Encoder-Decoder, GAN, Pix2Pix, RNN/LSTM and 3DMM	FSGAN [67], faceswap [66], Tacotron [233, 238], and Wave2Lip [111]
Avg. Duration	11.4 sec	4 sec	18 sec-	13 sec	30 sec	–	–	Varied	Varied
Resolution	294 × 500	64×64 (LQ) 128 ×128 (HQ)	480p, 720p, 1080p	Varied	180p – 2160p	1920 ×1080	Varied	Varied	225 ×224
Format	–	JPG	H.264, CRF=0, 23, 40	Mp4	H.264	Mp4	Mp4	Mp4	Mp4
Visual quality	Low	Low	Low	High	High	High	High	Both low and high	Low
Temporal flickering	Yes	Yes	Yes	Improved	Improved	Significantly improved	–	Significantly improved	Improved
Modality	Visual	Audio/visual	Visual	Visual	Audio/visual	Visual	Visual	Visual	Audio/visual

Table 18 Comparison of audio fakes detection datasets

	LJ speech dataset [303]	M-AILabs dataset [304]	Mozilla TTS [305]	FOR dataset [288]	Baidu Dataset [59]	ASV spoof 2019 [296]	WaveFake [297]
Released	2017	2019	2019	2019	2018	2019	2021
Total samples	13,100	–	5.5 million	195,000	120	122,157	117,985
Length (hrs)	24	999 hrs 32 min	7226	–	0.6	–	196 hrs
Speaker Accent	Native	Native	24% US English, 8% British English	Native	US English, British English	Native	Native
Languages	1	9	54	1	1	1	2
Speaker gender	100% Female	Male, female	47% Male 15% Female	50% male, 50% female	50% male, 50% female	43% male, 57 female	Female
Format	wav	wav	mp3	mp3	mp3	mp3	wav
Tool/ technology used for generation	recorded	recorded	recorded	Deep Voice 3, TTs, Google-Wavenet etc. [288]	Neural voice cloning [59]	Tacotron2 [238] and WaveNet [55]	GAN based TTS models

5.1 Video datasets

UADFV The first dataset released for deepfake detection was UADFV [74]. It consists of a total of 98 videos, where 49 are real videos collected from YouTube and then copies are manipulated by using the FakeApp application [42] to generate 49 fake videos. The average length of videos is 11.14 sec with an average resolution of 294 × 500 pixels. However, the visual quality of the videos is very low, and the resultant alteration is obvious and thus easy to detect.

DeepfakeTIMIT DeepfakeTIMIT [191] was introduced in 2018, and consists of a total of 620 videos of 32 subjects. For each subject, there are deepfake videos of two quality levels: DeepFake-TIMIT-LQ and DeepFake-TIMIT-HQ. In DeepFake-TIMIT-LQ, the resolution of the output image is 64 × 64, whereas in DeepFake-TIMIT-HQ, the resolution of output size is 128 × 128. The fake content is generated by employing a face swap-GAN [65]. The generated videos are only 4 seconds long, and the dataset contains no audio channel manipulation. Moreover, the resultant videos are often blurry and people in actual videos are mostly presented in full frontal face view with a monochrome color background.

FaceForensics++ One of the most famous datasets for deepfake detection is FF++ [95]. This dataset was presented in 2019 as an extended form of the FaceForensics dataset [306], which contains videos with facial expression manipulation only, and was released in 2018. The FF++ dataset has four subsets, named FaceSwap [307], DeepFake [43], Face2Face [38], and NeuralTextures [308]. The dataset contains 1000 original videos collected from the YouTube-8 M dataset [309] and 3000 manipulated videos generated using the computer graphics and deepfake approaches specified in [306]. This dataset is also available in two quality levels, uncompressed and H264 compressed format, which can be used to evaluate the performance of deepfake detection approaches on both compressed and uncompressed videos. The FF++ dataset fails to generalize lip-sync deepfakes however, and some videos exhibit color inconsistencies around the manipulated faces.

Celeb-DF Another popular dataset used for evaluating deepfake detection techniques is Celeb-DF [194]. This dataset presents videos of higher quality and tries to overcome the problem of visible source artifacts found in previous databases. The CelebDF dataset contains 408 original videos and 795 fake videos. The original content was collected from Youtube, and is divided into two parts named Real1 and Real2 respectively. In Real1, there are a total of 158 videos of 13 subjects with different gender and skin color. Real2 comprises 250 videos, each having a different subject, and the synthesized videos are generated from these original videos through the refinement of existing deepfake algorithms [310, 311].

Deepfake Detection Challenge (DFDC) Recently, the Facebook community launched a challenge, aptly named the Deepfake Detection Challenge (DFDC)-preview [312], and released a new dataset that contains 1131 original videos and 4119 manipulated videos. The altered content is generated using two unknown techniques. The final version of the DFDC database is publicly available on [298]. It contains 100,000 fake videos along with 19,000 original samples. The dataset is created using various face-swap-based methods with different augmentations, i.e., geometric and color transformations, varying frame rate, etc., and distractors, i.e., overlaying different types of objects, in a video.

DeeperForensics (DF) Another Large-Scale dataset for deepfake detection, containing 50,000 original and 10,000 manipulated videos, is found in [299]. A novel conditional autoencoder, namely DF-VAE, is used to create manipulated videos. The dataset comprises highly diverse samples in terms of actor appearance. Further, a mixture of distortions and perturbations, such as compression, blur, and noise, are added to better represent real-world scenarios. As compared to previous datasets [74, 191, 194], the quality of generated samples is significantly improved.

WildDeepfake WildDeepfake (WDF) [300] is considered to be one of the most challenging deepfake detection datasets. It contains both real and deepfake samples collected from the internet. This dataset contains video samples of diverse subject matter, along with variation in terms of resolution, background, illumination conditions, and compression rates.

ForgeryNet Another advanced Visual deepfakes-based dataset namely the ForgeryNet (FN) is presented in the ForgeryNet Challenge 2021 [301]. ForgeryNet is an extensive online available deep face forgery dataset comprising 2.9 million static samples, along with 221,247 videos. This dataset is created by applying different 7 image-level alteration techniques, and 8 video-level forgery methods. Furthermore, about 36 various perturbations attacks are added to make the dataset more challenging and close to real-world scenarios.

FakeAVCeleb FakeAVCeleb [302] dataset is recently released and contains multimodal deepfake videos that involve manipulation in both audio and video channels with accurate lip-syncing. The dataset is generated using real videos collected from YouTube and popular synthetic algorithms such as FSGAN [67], FaceSwap [66], Tacotron [233, 238], and Wave2Lip [111]. The dataset also includes fine-level video labelling respective to audio-visual manipulation, resulting

in four pair combinations: real audio-real video, real audio-fake video, fake audio-real video, fake audio-real video, and fake audio-fake video. Videos featuring celebrities from different ethnic backgrounds and ages, with equal representation of each gender, are included to eliminate racial biasness and improve the fairness of deepfake detectors.

A representative map for datasets based on release year and size is shown in Fig. 14. Furthermore, we have added the visual samples from the mentioned datasets to facilitate the reader to visually experience the synthesis quality of the DeepFake datasets in Fig. 15. All of the above-mentioned datasets contain synthesized face portions only; these datasets lack upper/full body deepfakes. A more robust dataset is needed which should be able to synthesize an entire body deepfake.

5.2 Audio datasets

LJ speech and M-AILabs dataset LJSpeech [303] and M-AILabs [304] datasets are famous for the real-speech database employed in numerous TTS applications, i.e. DeepVoice 3 [224]. The LJSpeech database is comprised of 13,100 clips totaling 24 hours in length. All samples are recorded by a female speaker. The M-AILABS dataset consists of total of 999 hours and 32 minutes of audio. This dataset was created with multiple speakers in 9 different languages.

Mozilla TTS Mozilla Firefox, a well-known publicly available browser, released the biggest open-source database of people speaking [305]. Initially, the database included 1400 hours of recorded voices, in 18 different languages, in 2019. Later it was extended to 7226 hours of recorded voices in 54 diverse languages. This dataset contains 5.5 million audio clips and was employed by Mozilla's Deep Speech toolkit.

ASV spoof 2019 Another well-known dataset for fake audio detection is ASVspoof-2019 [218], which is comprised of two parts for performing logical access (LA) and physical access (PA) state analysis. Both LA and PA are created from the VCTK base corpus, which comprises audio clips taken from 107 speakers (46 males, 61 females). LA consists of both voice cloning and voice conversion samples, whereas PA consists of replay samples along with bona fide ones. Both datasets are further divided into three databases, named training, development, and evaluation, which contain clips from 20- (8 males, 12 females), 10- (4 males, 6 females), and 48- (21 males, 27 females) speakers respectively. Further categorization is diverse in terms of presenters, and the recording situations are the same for all source samples. The training and development sets contain spoofing occurrences created with the same method/conditions (labeled as known attacks), while the evaluation set contains samples with unknown attacks.

Fake-or-Real (FOR) dataset The FOR database [288] is another dataset that is widely employed for synthetic voice detection. This database consists of over 195,000 samples both from humans and AI-synthetic speech. This database groups samples from the new TTS method (i.e. Deep Voice 3 [224] and Google-Wavenet [55]) together with diverse human speech samples (i.e. Arctic Dataset, LJSpeech Dataset, VoxForge Dataset). The FOR database has four versions, namely for-original (FO), for-norm (FN), for-2 sec (F2S), and for-rerec (FR). FO contains unbalanced voices without alteration, while FN comprises balanced unaltered samples in terms of gender, class, and volume, etc. F2S contains data from FN, however, the samples are trimmed to 2 seconds, and the FR version is a rerecorded version of the F2S database, to simulate a condition in which an attacker passes a sample via a voice channel (i.e. a cellphone call or a voice message).

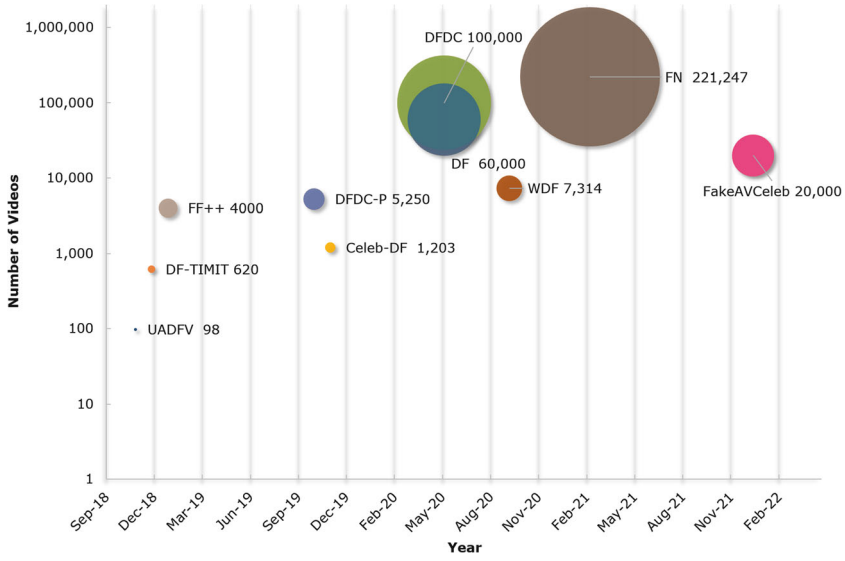

Fig. 14 Comparison of current video deepfake datasets over time based on the number of videos

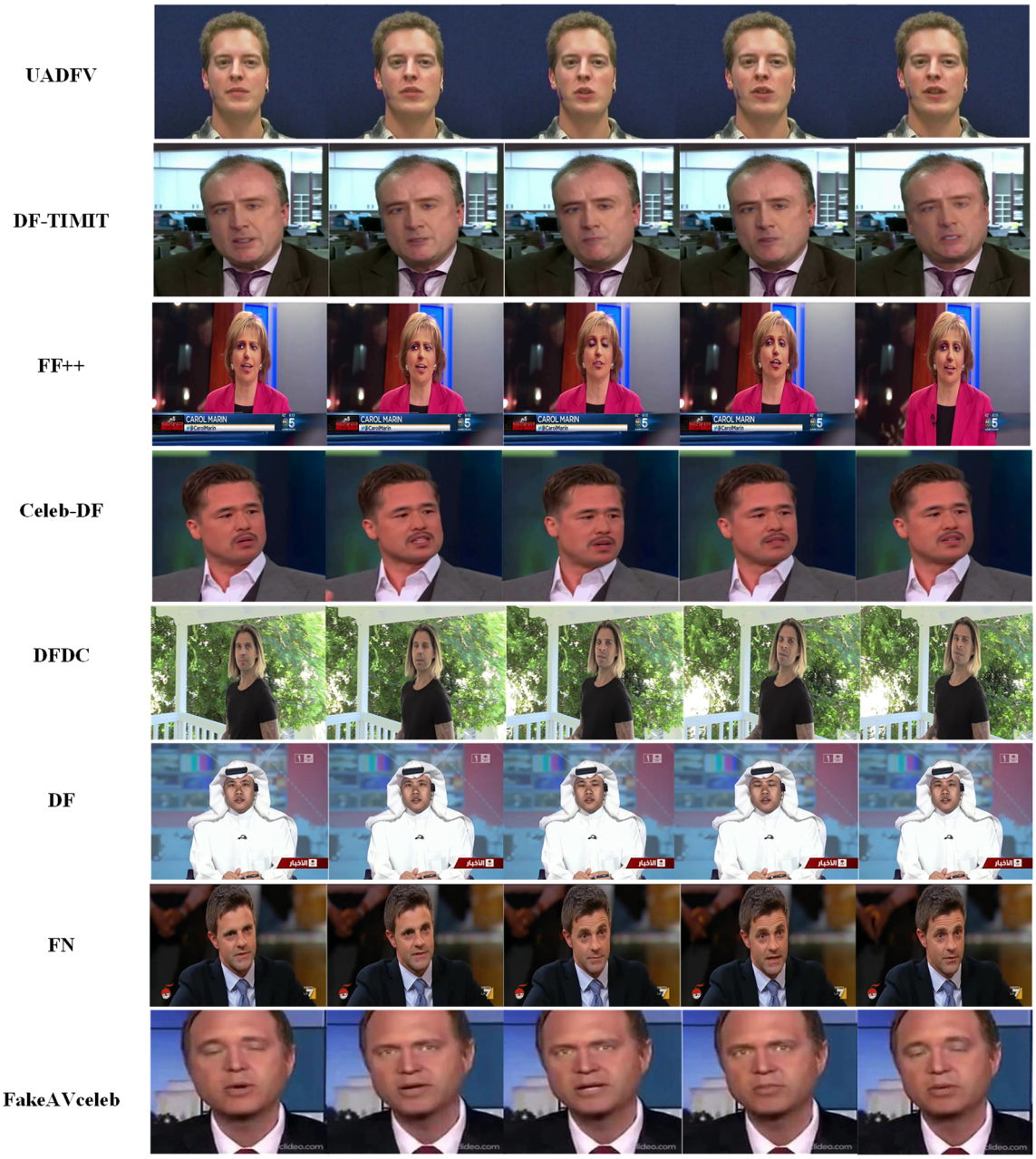

Fig. 15 Sample frames from different DeepFake video datasets

Baidu dataset The Baidu Silicon Valley AI Lab cloned audio dataset is another database employed for cloned speech detection [59]. This database is comprised of 10 ground truth speech recordings, 120 cloned samples, and 4 morphed samples.

ASV spoof 2021 The ASVspoof-2021 [296] is another dataset released as a part of ASVspoof challenge [276]. Along with earlier LA and PA partitions, this database includes an extra assessment partition for an audio deepfake detection track. This database is an extension of ASV spoof 2019 and has no specific training set. It includes only an evaluation set which comprises speech created from 48 speakers including 27 females and 21 males. This dataset is more challenging than previous versions and contains various audio coding and compression attacks with different environments and transmission scenarios.

WaveFake Recently, WaveFake (WF) [297] a large-scale audio deepfake detection dataset, was released. It contains 117,985 fake audio clips in 16-bit PCM wav format. The database uses six different advanced TTS audio generative models across two languages. The synthetic speech samples closely resemble real speech data,

however it lacks diversity and includes samples by only one speaker.

6 Open challenges

6.1 Open challenges in Deepfakes generation

Although extensive efforts have been shown to improve the visual quality of generated deepfakes there are still several challenges that need to be addressed. A few of them are discussed below.

Generalization The generative models are data-driven, and therefore they reflect the learned features during training, in the output. To generate high-quality deepfakes a large amount of data is required for training. Moreover, the training process itself requires hours to produce convincing deepfake audiovisual content. Usually, it is easier to obtain a dataset of the source/driving identity but the availability of sufficient data for a specific victim is a challenging task. Also, retraining the model for each specific target identity is computationally complex. Because of this, a generalized model is required to enable the execution of a trained model for multiple target identities unseen during training or with few training samples available.

Identity Leakage The preservation of target identity is a problem when there is a significant mismatch between the target identity and the source identity, specifically in face reenactment tasks where target expressions are driven by some source identity. The facial data of the source identity is partially transferred to the generated face. This occurs when training is performed on single or multiple identities, but data pairing is accomplished for the same identity.

Paired Training A trained, supervised model can generate high-quality output but at the expense of data pairing. Data pairing is concerned with generating the desired output by identifying similar input examples from the training data. This process is laborious and inapplicable to those scenarios where different facial behaviors and multiple identities are involved in the training stage.

Pose Variations and Distance from the camera Existing deepfake techniques generate good results of the target for frontal facial view. However, the quality of manipulated content degrades significantly for scenarios where a person is looking off-camera. This results in undesired visual artifacts around the facial region. Furthermore, another big challenge for convincing deepfake generation is the facial distance of the target from the camera, as an increase in distance from capturing devices results in low-quality face synthesis.

Illumination Conditions Current deepfake generation approaches produce fake information in a controlled environment with consistent lighting conditions. However, an abrupt change in illumination conditions such as in indoor/outdoor scenes results in color inconsistencies and strange artifacts in the resultant videos.

Occlusions One of the main challenges in deepfake generation is the occurrence of occlusion, which results when the face region of the source and victim are obscured with a hand, hair, glasses, or any other item. Moreover, occlusion can be the result of the hidden face or eye portion which eventually causes inconsistent facial features in the manipulated content.

Temporal Coherence Another drawback of generated deepfakes is the presence of evident artifacts like flickering and jittering among frames. These effects occur because the deepfake generation frameworks work on each frame without taking into account the temporal consistency. To overcome this limitation, some works either provide this context to generator or discriminator, consider temporal coherence losses, employ RNNs, or take a combination of all these approaches.

High-quality audio speech synthesis TTS and VC based on neural networks attempt to push the boundaries and generate realistic speech for real-world applications. Current generated audio speech signal, however, lack different artficacts that exist in human speech, such as pauses, varying emotions, realism, expressiveness, accent, robustness, and controllability. Several generative models, such as VAE [239, 262, 271], GAN [239, 255, 257], vocoders [228, 252], and end-to-end learning models [224, 238] are used to improve the quality of the synthesized audio signal. However, there is a need for improved modeling techniques that produce the speech which is spontaneous, expressive, and varies in style, to enhance the naturalness of generated audio samples.

Robust speech synthesis The synthesis of high-quality speech for different languages requires extensive training and labeled text data, and consumes huge computing resources. Such settings introduce an extensive computational burden which usually results in a tradeoff between the quality and inference time for generated audio content. The research community has taken several initiatives to introduce lightweight audio signal generation techniques, such as the ZeroSpeech Challenge [313], where speech signal is generated from audio data only. However, to cope with the real-world scenarios, there is a need for a more robust approach that can generate a high-quality signal from a small training dataset and low resource consumption.

Speech Adaptability The existing speech synthesis techniques are target-specific, i.e., they are capable of generating an audio

signal for the specific person on which the model is trained. Such approaches lack the ability to generate a high-quality signal for unseen instances, which shows a deficit in the ability to generalize in the existing speech synthesis models. The main reason models lack adaptability is over-fitting on training data, which makes them unable to efficiently learn enough acoustic information to be able to generate samples for a new target. Therefore, a more accurate, generalizable, model is required to tackle the current challenges of speech generation models [59, 235, 268].

Realism in synthetic audio speech Though the quality of synthetic audio is certainly getting much better, there is still a need for improvement. Some of the main challenges are lack of natural emotions, control over duration, sound volume, and the pace at which the target speaks. The existing speech generation models use one-to-many mappings [229, 240], which produce a low-quality speech signal with a lack of expressiveness in the presence of insufficient sample data. Therefore, there is a need for an efficient model that can better learn the varying qualities of speech signals in order to produce high-quality synthetic audio.

6.2 Challenges in deepfakes detection methods

Although remarkable advancements have been made in the performance of deepfake detectors, there are numerous concerns about current detection techniques that need attention. Some of the challenges of deepfake detection approaches are discussed in this section.

Quality of deepfake datasets The accessibility of large databases of deepfakes is an important factor in the generation of deepfake detection techniques. Analyzing the quality of videos from these datasets, however, reveals several ambiguities when compared to actual manipulated content found on the internet. Different visual artifacts that can be visualized in these databases are: i) temporal flickering in some cases during the speech, ii) blurriness around the facial regions, iii) over smoothness in facial texture/lack of facial texture details, iv) lack of head pose movement or rotation, v) lack of face occluding objects such as glasses, and lightning effects, vi) sensitivity to variations in input posture or gaze, skin color inconsistency, and identity leakage, and vii) limited availability of a combined high-quality audio-visual deepfake dataset. The aforementioned dataset ambiguities are due to imperfect steps in the manipulation technique. Furthermore, manipulated content of low quality can hardly be convincing or create a real impression. Therefore, even if detection approaches exhibit better performance over such videos it is not guaranteed that these methods will perform well when employed in the real-world scenarios.

Performance evaluation Presently, deepfake detection methods are formulated as a binary classification problem, where each sample can be either real or fake. Such classifiers are easier to build in a controlled environment, where we generate and verify deepfake detection techniques by utilizing audio-visual content that is either original or fabricated. However, for real-world scenarios, videos can be altered in ways other than deepfakes, so just because the content was not detected as manipulated does not guarantee the video is an original one. Furthermore, deepfake content can be the subject of multiple types of alteration, i.e., audio and/or visual, and therefore a single label may not be completely accurate. Moreover, in visual content with multiple people's faces, more than one of them could be manipulated with deepfakes over a segment of frames. Therefore, any binary classification scheme should be enhanced to multiclass/multi-label and utilize local classification/detection at the frame level to cope with the challenges of real-world scenarios.

Model scalability Another main challenge in the existing deepfake detection models is the lack of scalability for large-scale platforms, such as social media [197, 314]. When used in a real-world scenario, inference time becomes a critical factor for detecting fake audio-visual content. Designing a model with high accuracy but with a very long inference time makes the approach unlikely to be widely used in actual applications. Therefore, there is a need for detection techniques that have real-time performance capability with a high accuracy rate for massive deepfake content detection.

Explainability in detection methods Existing deepfake detection approaches are typically designed to perform batch analysis over a large dataset, however when these techniques are employed in the field by journalists or law enforcement, there may only be a small set of videos available for analysis. A numerical score parallel to the probability of an audio or video being real or fake is not as valuable to the practitioners if it cannot be confirmed with an appropriate proof of the score. In those situations, it is very common to demand an explanation for the numerical score for the analysis to be believed before publication or utilization in a court of law. Most deepfake detection methods lack such an explanation, however, particularly those which are based on DL approaches due to their black-box nature.

Fairness and trust It has been observed that existing audio and visual deepfakes datasets are biased and contain imbalanced data of different races and genders. Furthermore, the detection techniques employed can be biased as well. Although researchers have started doing work in this area to fill the gap very little work is available [315]. Hence, there is an urgent need to introduce approaches that improve the data and fairness in detection algorithms.

Temporal aggregation Existing deepfake detection methods are based on binary classification at the frame level, i.e. checking the probability that each video frame is real or manipulated. These approaches do not consider temporal consistency between frames, however, and suffer from two potential problems: (i) deepfake content shows temporal artifacts, and (ii) real or fake frames could appear in sequential intervals. Furthermore, these techniques require an extra step to compute the integrity score at the video level, as these methods need to combine the score from each frame to generate a final value.

Social media laundering Social platforms like Twitter, Facebook, or Instagram are among the main online networks used to spread audio-visual content among the general public. To save bandwidth on the network or to secure the user's privacy, such content is commonly stripped of meta-data, down-sampled, and substantially compressed before uploading. These manipulations, normally known as social media laundering, remove clues with respect to underlying forgeries and eventually increase false positive detection rates. Most deepfake detection approaches employ signal level keypoints and are more affected by social media laundering. A measure to increase the accuracy of deepfake identification approaches over social media laundering is to keenly include simulations of these effects in training data, and also increase the evaluation databases to contain data on social media laundered visual content.

Diversified audio DeepFake detection datasets Currently, extensive and diverse datasets for visual deepfake detection are available, however, there is a lack of such datasets for audio deepfake detection systems. Recently launched synthesized audio datasets, i.e., ASVspoof-2021 [296] and WaveFake [297] have been introduced, however the ASVspoof-2021 dataset does not contain specific training data for the audio deepfake track, while others contain samples from a single person only. Therefore, existing audio deepfakes detection approaches still require a more challenging and diverse dataset for the evaluation and detection of real-world deepfakes.

DeepFake detection evasion Most deepfake detection methods are concerned with missing information and artifacts left during the generation process. Detection techniques may fail, however, when this data is unavailable as attackers attempt to remove such traces during the manipulation generation process. Such fooling techniques are classified into three types: adversarial perturbation attacks, elimination of manipulation traces in the frequency domain, and the employment of image filtering to mislead detectors. In the case of visual adversarial attacks, different perturbations, such as random cropping, noise, and JPEG compression, are added to the training data, which ultimately results in high false alarms for detection methods. Different works [316, 317] have evaluated the performance of the state-of-the-art visual deepfake detectors in the presence of adversarial attacks and display an intense reduction in accuracy. In the case of audio, studies such as [318, 319] show that several adversarial pre/post-processing operations can be used to evade spoof detection. Similarly, the method in [320] is concerned with improving the quality of GAN-generated samples by enhancing spectral distributions. Such methods ultimately result in removing fake traces in the frequency domain and complicate the detection process [321, 322]. A third method, in [323–325], uses advanced image filtering techniques to improve generation quality such as the removal of model-based fingerprints left during generation and the addition of noise to remove fake signs. These methods pose a real challenge for deepfake detection methods, thus the research community needs to propose techniques that are robust and reliable to such attacks.

7 Future directions

Synthetic media is gaining a lot of attention because of its potential positive and negative impact on our society. The competition between deepfake generation and detection will not end in the foreseeable future, although impressive work has been presented for the generation and detection of deepfakes. There is still, however, room for improvement. In this section, we discuss the current state of deepfakes, their limitations, and future trends.

7.1 Creation

Visual media has more influence compared to text-based disinformation. Recently, the research community has focused more on the generation of identity agnostic models and high-quality deepfakes. A few distinguished improvements are i) a reduction in the amount of training data due to the introduction of un-paired self-supervised methods [326], ii) quick learning, which allows identity stealing using a single image [132, 134], iii) enhancements in visual details [60, 147], iv) improved temporal coherence in generated videos by employing optical flow estimation and GAN based temporal discriminators [107], v) the alleviation of visible artifacts around the face boundary by adding secondary networks for seamless blending [69], and vi) improvements in synthesized face quality by adding multiple losses with different responsibilities, such as occlusion, creation, conversion, and blending [112]. Several approaches have been proposed to boost the visual quality and realism of deepfake generation, however, there are a few limitations. Most of the current synthetic media generation focuses on a frontal face pose. In facial reenactment, for good results, the face is swapped with a lookalike

identity. However, it is not possible to always have the best match, which ultimately results in identity leakage.

AI-based manipulation is not restricted to the creation of visual content only, leading to a generation of highly genuine audio deepfakes. The quality of audio deepfakes has significantly improved and requires less training data to generate more realistic synthetic audio of the target speaker. The employment of synthesized speech for impersonating targets can produce highly convincing deepfakes with a marked negative adverse impact on society. Currently, audio-visual content is generated separately using multiple disconnected steps, which ultimately results in the generation of asynchronous content. Present deepfake generation focuses on the face region only, however the next generation of deepfakes is expected to target full body manipulations, such as a change in body pose, along with convincing expressions. Target-specific joint audio-visual synthesis with more naturalness and realism in speech is a new cutting-edge application of the technology in the context of persona appropriation [108, 327]. Another possible trend is the creation of real-time deepfakes. Some researchers have already reported attaining real-time deepfakes at 30fps [67]. Such alterations will result in the generation of more believable deepfakes.

7.2 Detection

To prevent deepfake misinformation and disinformation, some authors presented approaches to identify forensic changes made with visual content by employing the concept of blockchain and smart contracts [328–330]. In [329] the authors utilized Ethereum smart contracts to locate and track the origin and history of manipulated information and its source, even in the presence of multiple manipulation attacks. This smart contract applied hashes of the interplanetary file system to saved videos, together with their metadata. While this method may perform well for deepfake identification, it is applicable only if the video metadata exists. Thus, the development and adoption of such techniques could be useful for the newswires, however, the vast majority of content created by normal citizens won't be protected by such techniques.

Recent automated deepfake identification approaches typically deal with face swapping videos, and the majority of uploaded fake videos belong in this category. Major improvements in detection algorithms include i) identification of artifacts left during the generation process, such as inconsistencies in head pose [74], lack of eye blinking [80], color variations in facial texture [160] and teeth alignment, ii) detection of unseen GAN generated samples, iii) spatio-temporal features, and iv) physiological signals like heart rate [89], and an individual's behavior patterns [83]. Although extensive work has been presented for automated detection, these automated detection methods are expected to be short-lived and require improvements on multiple fronts. Following are many of unresolved challenges in the domain of deepfake detection.

- The existing methods are not robust to post-processing operations like compression, noisy effects, light variations, etc. Moreover, limited work has been presented that can detect both audio and visual deepfakes.
- Recently, most of the techniques have focused on face-swap detection by exploiting its limitations, like visible artifacts. However, with immense developments in technology, the near future will produce more sophisticated face-swaps, such as impersonating someone, with the target having a similar face shape, personality, and hairstyle. Aside from this, other types of deepfake, like face-reenactment and lip-synching are getting stronger day by day.
- The introduction of Vision Transformer techniques that use a self-attention mechanism to learn meaningful representation from the input has shown remarkable performance in a variety of machine vision tasks. The concept of patch embedding with CNN features can perform well for deepfake detection due to their accuracy and high recall rate. Even though some work has been presented by researchers [331–333] there is a need for more exploration of this concept as these approaches have the potential to better tackle the challenges of deepfake recognition, such as robustness against unseen manipulations and perturbation attacks.
- Existing deepfake detectors have mainly relied on the signatures of existing deepfakes by using ML techniques, including unsupervised clustering and supervised classification methods, and therefore are less likely to detect unknown deepfakes. Both anomaly-based and signature-based detection methods have their own pros and cons. For example, anomaly detection-based approaches show a high false alarm rate because they may misclassify a bona fide multimedia sample whose patterns are rare in the dataset. On the other hand, signature-based approaches cannot discover unknown attacks [334]. Therefore, a hybrid approach using both anomaly and signature-based detection needs to be studied in order to identify known and unknown attacks. Furthermore, a collaboration with the Reinforcement Learning (RL) method could be added to the hybrid signature and anomaly approach. More specifically, RL can give a reward (or penalty) to the system when it selects frames that contain (or do not contain) anomalies, or any signs of manipulation. Additionally, in the future, deep reinforcement active learning approaches [335, 336] could play a pivotal role in the detection of deepfakes.
- Anti-forensic, or adversarial, ML techniques can be employed to reduce the classification accuracy of automated detection methods. Game-theoretic approaches could be

employed to mitigate adversarial attacks on deepfake detectors. Additionally, RL, and particularly deep reinforcement learning (DRL), is extremely efficient in solving intricate cyber-defense problems. Thus, DRL could offer great potential for not only deepfake detection but also to counter antiforensic attacks on the detectors. Since RL can model an autonomous agent to take sequential actions optimally with limited, or without prior, knowledge of the environment, it could be used to meet a need for developing algorithms to capture traces of anti-forensic processing and to design attack-aware deepfake detectors. The defense of deepfake detectors against adversarial input could be modeled as a two-player zero-sum game with which player utilities sum to zero at each time step. The defender here is represented by an actor-critic DRL algorithm [337].

- The current deepfake detectors face challenges, particularly due to incomplete, sparse, and noisy data in the training phases. There is a need to explore innovative AI architectures, algorithms, and approaches that "bake in" physics, mathematics, and prior knowledge relevant to deepfakes. Embedding physics and prior knowledge using knowledge-infused learning into AI will help to overcome the challenges of sparse data and will facilitate the development of generative models that are causal and explanative.
- Most of the existing approaches have focused on one specific type of feature, such as landmark features. However, as the complexity of deepfakes is increasing, it is important to fuse landmarks, photoplethysmography (PPG), and audio-based features. Likewise, it is important to evaluate the fusion of classifiers. Particularly, the fusion of anomaly and signature-based ensemble learning will assist in the improvement of accuracy in deepfake detectors.
- Existing research on deepfakes has mainly focused on detecting manipulation in the visual content of the video, however, audio manipulation, an integral component of deepfakes, has been mostly ignored by the research community. There exists a need to develop unified deepfake detectors that are capable of effectively detecting both audio (i.e., TTS synthesis, voice conversion, cloned-replay) and visual forgeries (face-swap, lip-sync, and puppet-master) simultaneously.
- Existing deepfake datasets lack the potential attributes (i.e. multiple visual and audio forgeries, etc.) required to evaluate the performance of more robust deepfake detection methods. As stated above, the research community has hardly explored the fact that deepfake videos contain not only visual forgeries but audio manipulations as well. Existing deepfake datasets do not consider audio forgery and only focus on visual forgeries. In near future, the role of voice cloning (TTS synthesis, VC) and replay spoofing, may increase in deepfake video generation. Additionally, shallow audio forgeries can easily be fused along with deep audio forgeries in deepfake videos. We have already developed a voice spoofing detection corpus [338] for single- and multi-order replay attacks. Currently, we are working on developing a robust voice cloning and audio-visual deepfake dataset that can be effectively used to evaluate the performance of futuristic audio-visual deepfake detection methods.
- A unified method to address the variation of cloned attacks, such as cloned replay. The majority of voice spoofing detectors target detecting either replay or cloning attacks [218, 277, 286]. These two-class oriented, genuine vs. spoof countermeasures, are not ready to counter multiple spoofing attacks on automatic speaker verification (ASV) systems. A study on presentation attack detection indicated that the countermeasures trained on a specific type of spoofing attack do not generalize well for other types of spoofing attacks [339]. Moreover, a unified countermeasure that can detect replay and cloning attacks in multi-hop scenarios, where multiple microphones and smart speakers are chained together, does not exist. We addressed the problem of spoofing attack detection in multi-hop scenarios in our prior work [11], but only for voice replay attacks. Therefore, there exists an urgent need to develop a unified countermeasure that can effectively detect a variety of spoofing attacks (i.e. replay, cloning, and cloned replay) in a multi-hop scenario.
- The exponential growth of smart speakers and other voice-enabled devices has made Automated Speech Verification (ASV) a fundamental component. However, optimal utilization of ASV in critical domains, such as financial services, health care, etc., is not possible unless we counter the threats of multiple voice spoofing attacks on the ASV. Thus, this vulnerability also presents a need to develop a robust and unified spoofing countermeasure.
- There exists a crucial need to implement federated, learning-based, lightweight approaches to detect the manipulation at the source, so an attack doesn't traverse a network of smart speakers (or other IoT devices) [10, 11].

8 Conclusion

This survey paper presents a comprehensive review of existing deepfake generation and detection methods. Not all digital manipulations are harmful. Due to immense technological advancements, however, it is now very easy to produce realistic fabricated content. Therefore, malicious users can use it to spread disinformation, to attack individuals, and to cause social, psychological, religious, mental, and political stress. In the future, we imagine seeing the results of fabricated content in many other modalities and industries. There is a cold war

between deepfake generation and detection methods. As there are improvements in one it causes challenges for the other. We provided a detailed analysis of existing audio and video deepfake generation and detection techniques, along with their strengths and weaknesses. We have also discussed existing challenges and the future directions of both deepfake creation and identification methods.

Acknowledgements This material is based upon work supported by the National Science Foundation (NSF) under Grant number 1815724, Punjab Higher Education Commission of Pakistan under Award No. (PHEC/ARA/PIRCA/20527/21), and Michigan Translational Research and Commercialization (MTRAC) Advanced Computing Technologies (ACT) Grant Case number 292883. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF and MTRAC ACT.

References

1. Goodfellow I et al (2014) Generative adversarial nets. Adv Neural Inf Proces Syst 1:2672–2680
2. Etienne H (2021) The future of online trust (and why Deepfake is advancing it). AI Ethics 1:553–562. https://doi.org/10.1007/s43681-021-00072-1
3. ZAO. https://apps.apple.com/cn/app/zao/id1465199127. Accessed September 09, 2020
4. Reface App. https://reface.app/. Accessed September 11, 2020
5. FaceApp. https://www.faceapp.com/. Accessed September 17, 2020
6. Audacity. https://www.audacityteam.org/. Accessed September 09, 2020
7. Sound Forge. https://www.magix.com/gb/music/sound-forge/. Accessed January 11, 2021
8. Shu K, Wang S, Lee D, Liu H (2020) Mining disinformation and fake news: concepts, methods, and recent advancements. In: Disinformation, misinformation, and fake news in social media. Springer, pp 1–19
9. 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
10. Malik KM, Malik H, Baumann R (2019) Towards vulnerability analysis of voice-driven interfaces and countermeasures for replay attacks. In 2019 IEEE conference on multimedia information processing and retrieval (MIPR). IEEE, pp 523–528
11. Malik KM, Javed A, Malik H, Irtaza A (2020) A light-weight replay detection framework for voice controlled iot devices. IEEE J Sel Top Sign Process 14:982–996
12. Javed A, Malik KM, Irtaza A, Malik H (2021) Towards protecting cyber-physical and IoT systems from single-and multi-order voice spoofing attacks. Appl Acoust 183:108283
13. Aljasem M, Irtaza A, Malik H, Saba N, Javed A, Malik KM, Meharmohammadi M (2021) Secure automatic speaker verification (SASV) system through sm-ALTP features and asymmetric bagging. IEEE Trans Inf Forensics Secur 16:3524–3537
14. Sharma M, Kaur M (2022) A review of Deepfake technology: an emerging AI threat. Soft Comput Secur Appl:605–619
15. Zhang T (2022) Deepfake generation and detection, a survey. Multimed Tools Appl 81:6259–6276. https://doi.org/10.1007/s11042-021-11733-y
16. Malik A, Kuribayashi M, Abdullahi SM, Khan AN (2022) DeepFake detection for human face images and videos: a survey. IEEE Access 10:18757–18775
17. Rana MS, Nobi MN, Murali B, Sung AH (2022) Deepfake detection: a systematic literature review. IEEE Access
18. Verdoliva L (2020) Media forensics and deepfakes: an overview. IEEE J Sel Top Sign Process 14:910–932
19. Tolosana R, Vera-Rodriguez R, Fierrez J, Morales A, Ortega-Garcia J (2020) Deepfakes and beyond: a survey of face manipulation and fake detection. Inf Fusion 64:131–148
20. Nguyen TT, Nguyen CM, Nguyen DT, Nguyen DT, Nahavandi S (2019) Deep learning for deepfakes creation and detection. arXiv preprint arXiv:190911573
21. Mirsky Y, Lee W (2021) The creation and detection of deepfakes: a survey. ACM Comput Surv 54:1–41
22. Oliveira L (2017) The current state of fake news. Procedia Comput Sci 121:817–825
23. Chesney R, Citron D (2019) Deepfakes and the new disinformation war: the coming age of post-truth geopolitics. Foreign Aff 98:147
24. Karnouskos S (2020) Artificial intelligence in digital media: the era of deepfakes. IEEE Trans Technol Soc 1:138–147
25. Stiff H, Johansson F (2021) Detecting computer-generated disinformation. Int J Data Sci Anal 13:363–383. https://doi.org/10.1007/s41060-021-00299-5
26. Dobber T, Metoui N, Trilling D, Helberger N, de Vreese C (2021) Do (microtargeted) deepfakes have real effects on political attitudes? Int J Press Polit 26:69–91
27. Lingam G, Rout RR, Somayajulu DV (2019) Adaptive deep Q-learning model for detecting social bots and influential users in online social networks. Appl Intell 49:3947–3964
28. Shao C, Ciampaglia GL, Varol O, Yang K-C, Flammini A, Menczer F (2018) The spread of low-credibility content by social bots. Nat Commun 9:1–9
29. Marwick A, Lewis R (2017) Media manipulation and disinformation online. Data & Society Research Institute, New York, pp 7–19
30. Tsao S-F, Chen H, Tisseverasinghe T, Yang Y, Li L, Butt ZA (2021) What social media told us in the time of COVID-19: a scoping review. Lancet Digit Health 3:e175–e194
31. Pierri F, Ceri S (2019) False news on social media: a data-driven survey. ACM SIGMOD Rec 48:18–27
32. Chesney B, Citron D (2019) Deep fakes: a looming challenge for privacy, democracy, and national security. Calif Law Rev 107:1753
33. 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
34. Gupta S, Mohan N, Kaushal P (2021) Passive image forensics using universal techniques: a review. Artif Intell Rev 1:1–51
35. Pavan Kumar MR, Jayagopal P (2021) Generative adversarial networks: a survey on applications and challenges. Int J Multimed Inf Retr 10:1–24. https://doi.org/10.1007/s13735-020-00196-w
36. Choi Y, Choi M, Kim M, Ha J-W, 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
37. Suwajanakorn S, Seitz SM, Kemelmacher-Shlizerman I (2017) Synthesizing Obama: learning lip sync from audio. ACM Trans Graph 36:95–108. https://doi.org/10.1145/3072959.3073640
38. 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

39. Wiles O, Sophia Koepke A, Zisserman A (2018) X2face: a network for controlling face generation using images, audio, and pose codes. In: Proceedings of the European conference on computer vision (ECCV), pp 670–686
40. 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
41. Johnson DG, Diakopoulos N (2021) What to do about deepfakes. Commun ACM 64:33–35
42. FakeApp 2.2.0. https://www.malavida.com/en/soft/fakeapp/. Accessed September 18, 2020
43. Faceswap: Deepfakes software for all. https://github.com/deepfakes/faceswap. Accessed September 08, 2020
44. DeepFaceLab. https://github.com/iperov/DeepFaceLab. Accessed August 18, 2020
45. Siarohin A, Lathuilière S, Tulyakov S, Ricci E, Sebe N (2019) First order motion model for image animation. In: Advances in neural information processing systems, pp 7137–7147
46. Zhou H, Sun Y, Wu W, Loy CC, Wang X, Liu Z (2021) Pose-controllable talking face generation by implicitly modularized audio-visual representation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 4176–4186
47. Kim H, Garrido P, Tewari A, Xu W, Thies J, Niessner M, Pérez P, Richardt C, Zollhöfer M, Theobalt C (2018) Deep video portraits. ACM Trans Graph 37:163–177. https://doi.org/10.1145/3197517.3201283
48. Ha S, Kersner M, Kim B, Seo S, Kim D (2020) Marionette: few-shot face reenactment preserving identity of unseen targets. In: Proceedings of the AAAI conference on artificial intelligence, pp 10893–10900
49. Wang Y, Bilinski P, Bremond F, Dantcheva A (2020) ImaGINator: conditional Spatio-temporal GAN for video generation. In: The IEEE winter conference on applications of computer vision, pp 1160–1169
50. Lu Y, Chai J, Cao X (2021) Live speech portraits: real-time photorealistic talking-head animation. ACM Trans Graph 40:1–17
51. Lahiri A, Kwatra V, Frueh C, Lewis J, Bregler C (2021) LipSync3D: data-efficient learning of personalized 3D talking faces from video using pose and lighting normalization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2755–2764
52. Westerlund M (2019) The emergence of deepfake technology: a review. Technol Innov Manag Rev 9:39–52
53. Greengard S (2019) Will deepfakes do deep damage? Commun ACM 63:17–19
54. Lee Y, Huang K-T, Blom R, Schriner R, Ciccarelli CA (2021) To believe or not to believe: framing analysis of content and audience response of top 10 deepfake videos on youtube. Cyberpsychol Behav Soc Netw 24:153–158
55. Oord Avd et al. (2016) Wavenet: a generative model for raw audio. In: 9th ISCA speech synthesis workshop, p 2
56. Wang Y et al. (2017) Tacotron: towards end-to-end speech synthesis. arXiv preprint arXiv:170310135
57. Arik SO et al. (2017) Deep voice: real-time neural text-to-speech. In: International conference on machine learning PMLR, pp 195–204
58. 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. In: Proceedings of the 28th ACM international conference on multimedia, pp 1207–1216
59. Arik S, Chen J, Peng K, Ping W, Zhou Y (2018) Neural voice cloning with a few samples. In: Advances in neural information processing systems, pp 10019–10029
60. Wang T-C, Liu M-Y, Zhu J-Y, Tao A, Kautz J, Catanzaro B (2018) High-resolution image synthesis and semantic manipulation with conditional gans. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8798–8807
61. Nirkin Y, Masi I, Tuan AT, Hassner T, Medioni G (2018) On face segmentation, face swapping, and face perception. In 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018). IEEE, pp 98–105
62. Bitouk D, Kumar N, Dhillon S, Belhumeur P, Nayar SK (2008) Face swapping: automatically replacing faces in photographs. In: ACM transactions on graphics (TOG). ACM, pp 39
63. Lin Y, Lin Q, Tang F, Wang S (2012) Face replacement with large-pose differences. In: Proceedings of the 20th ACM international conference on multimedia. ACM, pp 1249–1250
64. Smith BM, Zhang L (2012) Joint face alignment with non-parametric shape models. In: European conference on computer vision. Springer, pp 43–56
65. Faceswap-GAN https://github.com/shaoanlu/faceswap-GAN. Accessed September 18, 2020
66. Korshunova I, Shi W, Dambre J, Theis L (2017) Fast face-swap using convolutional neural networks. In: Proceedings of the IEEE international conference on computer vision, pp 3677–3685
67. Nirkin Y, Keller Y, Hassner T (2019) FSGAN: subject agnostic face swapping and reenactment. In: Proceedings of the IEEE international conference on computer vision, pp 7184–7193
68. Natsume R, Yatagawa T, Morishima S (2018) RSGAN: face swapping and editing using face and hair representation in latent spaces. arXiv preprint arXiv:180403447
69. Natsume R, Yatagawa T, Morishima S (2018) Fsnet: an identity-aware generative model for image-based face swapping. In: Asian conference on computer vision. Springer, pp 117–132
70. Li L, Bao J, Yang H, Chen D, Wen F (2020) Advancing high fidelity identity swapping for forgery detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 5074–5083
71. Petrov I et al. (2020) DeepFaceLab: a simple, flexible and extensible face swapping framework. arXiv preprint arXiv:200505535
72. Chen D, Chen Q, Wu J, Yu X, Jia T (2019) Face swapping: realistic image synthesis based on facial landmarks alignment. Math Probl Eng 2019
73. Zhang Y, Zheng L, Thing VL (2017) Automated face swapping and its detection. In: 2017 IEEE 2nd international conference on signal and image processing (ICSIP). IEEE, pp 15–19
74. Yang X, Li Y, Lyu S (2019) Exposing deep fakes using inconsistent head poses. In: 2019 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 8261–8265
75. Güera D, Baireddy S, Bestagini P, Tubaro S, Delp EJ (2019) We need no pixels: video manipulation detection using stream descriptors. arXiv preprint arXiv:190608743
76. Jack K (2011) Video demystified: a handbook for the digital engineer. Elsevier
77. Ciftci UA, Demir I (2020) FakeCatcher: detection of synthetic portrait videos using biological signals. IEEE Trans Pattern Anal Mach Intell 1
78. Jung T, Kim S, Kim K (2020) DeepVision: Deepfakes detection using human eye blinking pattern. IEEE Access 8:83144–83154
79. Ranjan R, Patel VM, Chellappa R (2017) Hyperface: a deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. IEEE Trans Pattern Anal Mach Intell 41:121–135
80. Soukupova T, Cech J (2016) Eye blink detection using facial landmarks. In: 21st Computer Vision Winter Workshop
81. 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
82. Malik J, Belongie S, Leung T, Shi J (2001) Contour and texture analysis for image segmentation. Int J Comput Vis 43:7–27
83. 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
84. Li Y, Lyu S (2019) Exposing deepfake videos by detecting face warping artifacts. In: IEEE conference on computer vision and pattern recognition workshops (CVPRW), pp 46–52
85. Li Y, Chang M-C, Lyu S (2018) In ictu oculi: exposing ai generated fake face videos by detecting eye blinking. In: 2018 IEEE international workshop on information forensics and security (WIFS). IEEE, pp 1–7
86. Montserrat DM et al. (2020) Deepfakes detection with automatic face weighting. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, pp 668–669
87. de Lima O, Franklin S, Basu S, Karwoski B, George A (2020) Deepfake detection using spatiotemporal convolutional networks. arXiv preprint arXiv:14749
88. Agarwal S, El-Gaaly T, Farid H, Lim S-N (2020) Detecting deepfake videos from appearance and behavior. In 2020 IEEE international workshop on information forensics and security (WIFS). IEEE, pp 1–6
89. 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
90. Yang J, Xiao S, Li A, Lu W, Gao X, Li Y (2021) MSTA-net: forgery detection by generating manipulation trace based on multi-scale self-texture attention. IEEE Trans Circuits Syst Video Technol
91. Sabir E, Cheng J, Jaiswal A, AbdAlmageed W, Masi I, Natarajan P (2019) Recurrent convolutional strategies for face manipulation detection in videos. Interfaces (GUI) 3:80–87
92. 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
93. Nguyen HH, Fang F, Yamagishi J, Echizen I (2019) Multi-task learning for detecting and segmenting manipulated facial images and videos. In: 2019 IEEE 10th international conference on biometrics theory, applications and systems (BTAS), pp 1–8
94. Cozzolino D, Thies J, Rössler A, Riess C, Nießner M, Verdoliva L (2018) Forensictransfer: weakly-supervised domain adaptation for forgery detection. arXiv preprint arXiv:181202510
95. Rossler A, Cozzolino D, Verdoliva L, Riess C, Thies J, Nießner M (2019) Faceforensics++: learning to detect manipulated facial images. In: Proceedings of the IEEE international conference on computer vision, pp 1–11
96. King DE (2009) Dlib-ml: a machine learning toolkit. J Mach Learn Res 10:1755–1758
97. Zhang K, Zhang Z, Li Z, Qiao Y (2016) Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process Lett 23:1499–1503
98. Wiles O, Koepke A, Zisserman A (2018) Self-supervised learning of a facial attribute embedding from video. Paper presented at the 29th British machine vision conference (BMVC)
99. Rezende DJ, Mohamed S, Wierstra D (2014) Stochastic backpropagation and approximate inference in deep generative models. Paper presented at the international conference on machine learning, pp 1278–1286
100. Rahman H, Ahmed MU, Begum S, Funk P (2016) Real time heart rate monitoring from facial RGB color video using webcam. In: The 29th annual workshop of the Swedish artificial intelligence society (SAIS). Linköping University Electronic Press
101. Wu H-Y, Rubinstein M, Shih E, Guttag J, Durand F, Freeman W (2012) Eulerian video magnification for revealing subtle changes in the world. ACM Trans Graph 31:1–8
102. Chen RT, Rubanova Y, Bettencourt J, Duvenaud DK (2018) Neural ordinary differential equations. In: Advances in neural information processing systems, pp 6571–6583
103. Yang J, Li A, Xiao S, Lu W, Gao X (2021) MTD-net: learning to detect deepfakes images by multi-scale texture difference. IEEE Trans Inf Forensics Secur 16:4234–4245
104. Fan B, Wang L, Soong FK, Xie L (2015) Photo-real talking head with deep bidirectional LSTM. In: 2015 IEEE international conference on acoustics, Speech and Signal Processing (ICASSP). IEEE, pp 4884–4888
105. Charles J, Magee D, Hogg D (2016) Virtual immortality: reanimating characters from tv shows. In European conference on computer vision. Springer, pp 879–886
106. Jamaludin A, Chung JS, Zisserman A (2019) You said that?: Synthesising talking faces from audio. Int J Comput Vis 1:1–13
107. Vougioukas K, Petridis S, Pantic M (2019) End-to-end speech-driven realistic facial animation with temporal GANs. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 37–40
108. Zhou H, Liu Y, Liu Z, Luo P, Wang X (2019) Talking face generation by adversarially disentangled audio-visual representation. In: Proceedings of the AAAI conference on artificial intelligence, pp 9299–9306
109. 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. Wiley Online Library, pp 193–204
110. KR Prajwal, Mukhopadhyay R, Philip J, Jha A, Namboodiri V, Jawahar C (2019) Towards automatic face-to-face translation. In: Proceedings of the 27th ACM international conference on multimedia, pp 1428–1436
111. Prajwal K, Mukhopadhyay R, Namboodiri VP, Jawahar C (2020) A lip sync expert is all you need for speech to lip generation in the wild. In: Proceedings of the 28th ACM international conference on multimedia, pp 484–492
112. Fried O, Tewari A, Zollhöfer M, Finkelstein A, Shechtman E, Goldman DB, Genova K, Jin Z, Theobalt C, Agrawala M (2019) Text-based editing of talking-head video. ACM Trans Graph 38:1–14
113. Kim B-H, Ganapathi V (2019) LumiereNet: lecture video synthesis from audio. arXiv preprint arXiv:190702253
114. Korshunov P, Marcel S (2018) Speaker inconsistency detection in tampered video. In 2018 26th European signal processing conference (EUSIPCO). IEEE, pp 2375–2379
115. Sanderson C, Lovell BC (2009) Multi-region probabilistic histograms for robust and scalable identity inference. In: International conference on biometrics. Springer, pp 199–208
116. Anand A, Labati RD, Genovese A, Muñoz E, Piuri V, Scotti F (2017) Age estimation based on face images and pre-trained convolutional neural networks. In: 2017 IEEE symposium series on computational intelligence (SSCI). IEEE, pp 1–7
117. Boutellaa E, Boulkenafet Z, Komulainen J, Hadid A (2016) Audiovisual synchrony assessment for replay attack detection in talking face biometrics. Multimed Tools Appl 75:5329–5343
118. Korshunov P et al. (2019) Tampered speaker inconsistency detection with phonetically aware audio-visual features. In: International Conference on Machine Learning
119. Agarwal S, Farid H, Fried O, Agrawala M (2020) Detecting deepfake videos from phoneme-viseme mismatches. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, pp 660–661

120. Haliassos A, Vougioukas K, Petridis S, Pantic M (2021) Lips Don't lie: a Generalisable and robust approach to face forgery detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 5039–5049
121. Chugh K, Gupta P, Dhall A, Subramanian R (2020) Not made for each other-audio-visual dissonance-based deepfake detection and localization. In: Proceedings of the 28th ACM international conference on multimedia, pp 439–447
122. Mittal T, Bhattacharya U, Chandra R, Bera A, Manocha D (2020) Emotions Don't lie: an audio-visual deepfake detection method using affective cues. In: Proceedings of the 28th ACM international conference on multimedia, pp 2823–2832
123. 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 Sign Process 14:1024–1037
124. Thies J, Zollhöfer M, Theobalt C, Stamminger M, Nießner M (2018) Real-time reenactment of human portrait videos. ACM Trans Graph 37:1–13. https://doi.org/10.1145/3197517.3201350
125. Thies J, Zollhöfer M, Nießner M, Valgaerts L, Stamminger M, Theobalt C (2015) Real-time expression transfer for facial reenactment. ACM Trans Graph 34:1–14
126. Zollhöfer M, Nießner M, Izadi S, Rehmann C, Zach C, Fisher M, Wu C, Fitzgibbon A, Loop C, Theobalt C, Stamminger M (2014) Real-time non-rigid reconstruction using an RGB-D camera. ACM Trans Graph 33:1–12
127. Thies J, Zollhöfer M, Theobalt C, Stamminger M, Nießner M (2018) Headon: real-time reenactment of human portrait videos. ACM Trans Graph 37:1–13
128. Mirza M, Osindero S (2014) Conditional generative adversarial nets. arXiv preprint arXiv:14111784
129. Wu W, Zhang Y, Li C, Qian C, Change Loy C (2018) ReenactGAN: learning to reenact faces via boundary transfer. In: Proceedings of the European conference on computer vision (ECCV), pp 603–619
130. Pumarola A, Agudo A, Martínez AM, Sanfeliu A, Moreno-Noguer F (2018) GANimation: anatomically-aware facial animation from a single image. In: Proceedings of the European conference on computer vision (ECCV), pp 818–833
131. Sanchez E, Valstar M (2020) Triple consistency loss for pairing distributions in GAN-based face synthesis. In: 15th IEEE international conference on automatic face and gesture recognition. IEEE, pp 53–60
132. Zakharov E, Shysheya A, Burkov E, Lempitsky V (2019) Few-shot adversarial learning of realistic neural talking head models. In: Proceedings of the IEEE international conference on computer vision, pp 9459–9468
133. Zhang Y, Zhang S, He Y, Li C, Loy CC, Liu Z (2019) One-shot face reenactment. Paper presented at the British machine vision conference (BMVC)
134. Hao H, Baireddy S, Reibman AR, Delp EJ (2020) FaR-GAN for one-shot face reenactment. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR)
135. Blanz V, Vetter T (1999) A morphable model for the synthesis of 3D faces. In: Proceedings of the 26th annual conference on Computer graphics and interactive techniques, pp 187–194
136. Wehrbein T, Rudolph M, Rosenhahn B, Wandt B (2021) Probabilistic monocular 3d human pose estimation with normalizing flows. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 11199–11208
137. Lorenzo-Trueba J, Yamagishi J, Toda T, Saito D, Villavicencio F, Kinnunen T, Ling Z (2018) The voice conversion challenge 2018: promoting development of parallel and nonparallel methods. In the speaker and language recognition workshop. ISCA, pp 195–202
138. 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
139. Alparone L, Barni M, Bartolini F, Caldelli R (1999) Regularization of optic flow estimates by means of weighted vector median filtering. IEEE Trans Image Process 8:1462–1467
140. Sun D, Yang X, Liu M-Y, Kautz J (2018) PWC-net: CNNs for optical flow using pyramid, warping, and cost volume. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8934–8943
141. Baltrušaitis T, Robinson P, Morency L-P (2016) Openface: an open source facial behavior analysis toolkit. In: 2016 IEEE winter conference on applications of computer vision (WACV). IEEE, pp 1–10
142. Kingma DP, Welling M (2013) Auto-encoding variational bayes. arXiv preprint arXiv:13126114
143. Radford A, Metz L, Chintala S (2015) Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:151106434
144. Liu M-Y, Tuzel O (2016) Coupled generative adversarial networks. In: Advances in neural information processing systems, pp 469–477
145. Karras T, Aila T, Laine S, Lehtinen J (2017) Progressive growing of gans for improved quality, stability, and variation. In: 6th International Conference on Learning Representations
146. 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
147. Karras T, Laine S, Aittala M, Hellsten J, Lehtinen J, Aila T (2020) Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 8110–8119
148. Huang R, Zhang S, Li T, He R (2017) Beyond face rotation: global and local perception Gan for photorealistic and identity preserving frontal view synthesis. In: Proceedings of the IEEE international conference on computer vision, pp 2439–2448
149. Zhang H, Goodfellow I, Metaxas D, Odena A (2019) Self-attention generative adversarial networks. In: international conference on machine learning. PMLR, pp 7354–7363
150. Brock A, Donahue J, Simonyan K (2019) Large scale gan training for high fidelity natural image synthesis. In: 7th International Conference on Learning Representations
151. Zhang H, Xu T, Li H, Zhang S, Wang X, Huang X, Metaxas DN (2017) Stackgan: text to photo-realistic image synthesis with stacked generative adversarial networks. In: Proceedings of the IEEE international conference on computer vision, pp 5907–5915
152. Lu E, Hu X (2022) Image super-resolution via channel attention and spatial attention. Appl Intell 52:2260–2268. https://doi.org/10.1007/s10489-021-02464-6
153. Zhong J-L, Pun C-M, Gan Y-F (2020) Dense moment feature index and best match algorithms for video copy-move forgery detection. Inf Sci 537:184–202
154. Ding X, Huang Y, Li Y, He J (2020) Forgery detection of motion compensation interpolated frames based on discontinuity of optical flow. Multimed Tools Appl:1–26
155. Niyishaka P, Bhagvati C (2020) Copy-move forgery detection using image blobs and BRISK feature. Multimed Tools Appl:1–15
156. Sunitha K, Krishna A, Prasad B (2022) Copy-move tampering detection using keypoint based hybrid feature extraction and improved transformation model. Appl Intell:1–12
157. Tyagi S, Yadav D (2022) A detailed analysis of image and video forgery detection techniques. Vis Comput:1–21

158. Nawaz M, Mehmood Z, Nazir T, Masood M, Tariq U, Mahdi Munshi A, Mehmood A, Rashid M (2021) Image authenticity detection using DWT and circular block-based LTrP features. Comput Mater Contin 69:1927–1944
159. Akhtar Z, Dasgupta D (2019) A comparative evaluation of local feature descriptors for deepfakes detection. In: 2019 IEEE international symposium on technologies for homeland security (HST). IEEE, pp 1–5
160. McCloskey S, Albright M (2018) Detecting gan-generated imagery using color cues. arXiv preprint arXiv:08247
161. 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
162. Nataraj L, Mohammed TM, Manjunath B, Chandrasekaran S, Flenner A, Bappy JH, Roy-Chowdhury AK (2019) Detecting GAN generated fake images using co-occurrence matrices. Electronic Imaging 5:532–531
163. Yu N, Davis LS, Fritz M (2019) Attributing fake images to GANs: learning and analyzing GAN fingerprints. In: Proceedings of the IEEE international conference on computer vision, pp 7556–7566
164. Marra F, Saltori C, Boato G, Verdoliva L (2019) Incremental learning for the detection and classification of GAN-generated images. In: 2019 IEEE international workshop on information forensics and security (WIFS). IEEE, pp 1–6
165. Rebuffi S-A, Kolesnikov A, Sperl G, Lampert CH (2017) ICARL: incremental classifier and representation learning. In: proceedings of the IEEE conference on computer vision and pattern recognition, pp 2001–2010
166. Perarnau G, Van De Weijer J, Raducanu B, Álvarez JM (2016) Invertible conditional gans for image editing. arXiv preprint arXiv:161106355
167. Lample G, Zeghidour N, Usunier N, Bordes A, Denoyer L, Ranzato MA (2017) Fader networks: manipulating images by sliding attributes. In: Advances in neural information processing systems, pp 5967–5976
168. Choi Y, Uh Y, Yoo J, Ha J-W (2020) Stargan v2: diverse image synthesis for multiple domains. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 8188–8197
169. 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 28:5464–5478
170. 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
171. Zhang G, Kan M, Shan S, Chen X (2018) Generative adversarial network with spatial attention for face attribute editing. In: Proceedings of the European conference on computer vision (ECCV), pp 417–432
172. He Z, Kan M, Zhang J, Shan S (2020) PA-GAN: progressive attention generative adversarial network for facial attribute editing. arXiv preprint arXiv:200705892
173. Nataraj L, Mohammed TM, Manjunath B, Chandrasekaran S, Flenner A, Bappy JH, Roy-Chowdhury AK (2019) Detecting GAN generated fake images using co-occurrence matrices. Electron Imaging 2019:532-531–532-537
174. Zhang X, Karaman S, Chang S-F (2019) Detecting and simulating artifacts in gan fake images. In 2019 IEEE international workshop on information forensics and security (WIFS). IEEE, pp 1–6
175. Isola P, Zhu J-Y, Zhou T, Efros AA (2017) Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1125–1134
176. Wang R, Juefei-Xu F, Ma L, Xie X, Huang Y, Wang J, Liu Y (2021) Fakespotter: a simple yet robust baseline for spotting AI-synthesized fake faces. In: Proceedings of the 29th international conference on international joint conferences on artificial intelligence, pp 3444–3451
177. Parkhi OM, Vedaldi A, Zisserman A (2015) Deep face recognition. In: Proceedings of the British Machine Vision, pp 6
178. Amos B, Ludwiczuk B, Satyanarayanan M (2016) Openface: a general-purpose face recognition library with mobile applications. CMU School of Computer Science 6
179. Schroff F, Kalenichenko D, Philbin J (2015) Facenet: a unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 815–823
180. Bharati A, Singh R, Vatsa M, Bowyer KW (2016) Detecting facial retouching using supervised deep learning. IEEE Trans Inf Forensics Secur 11:1903–1913
181. Jain A, Singh R, Vatsa M (2018) On detecting gans and retouching based synthetic alterations. In: 2018 IEEE 9th international conference on biometrics theory, applications and systems (BTAS). IEEE, pp 1–7
182. Tariq S, Lee S, Kim H, Shin Y, Woo SS (2018) Detecting both machine and human created fake face images in the wild. In: Proceedings of the 2nd international workshop on multimedia privacy and security, pp 81–87
183. Dang H, Liu F, Stehouwer J, Liu X, Jain AK (2020) On the detection of digital face manipulation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 5781–5790
184. Rathgeb C, Botaljov A, Stockhardt F, Isadskiy S, Debiasi L, Uhl A, Busch C (2020) PRNU-based detection of facial retouching. IET Biom 9:154–164
185. Li Y, Zhang C, Sun P, Ke L, Ju Y, Qi H, Lyu S (2021) DeepFake-o-meter: an open platform for DeepFake detection. In: 2021 IEEE security and privacy workshops (SPW). IEEE, pp 277–281
186. Mehta V, Gupta P, Subramanian R, Dhall A (2021) FakeBuster: a DeepFakes detection tool for video conferencing scenarios. In 26th international conference on intelligent user interfaces, pp 61–63
187. Reality Defender 2020: A FORCE AGAINST DEEPFAKES. (2020). https://rd2020.org/index.html. Accessed August 03, 2021
188. Durall R, Keuper M, Pfreundt F-J, Keuper J (2019) Unmasking deepfakes with simple features. arXiv preprint arXiv:00686
189. Marra F, Gragnaniello D, Cozzolino D, Verdoliva L (2018) Detection of gan-generated fake images over social networks. In: 2018 IEEE conference on multimedia information processing and retrieval (MIPR). IEEE, pp 384–389
190. 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
191. Korshunov P, Marcel S (2018) Deepfakes: a new threat to face recognition? Assessment and detection. arXiv preprint arXiv:181208685
192. Wang S-Y, Wang O, Zhang R, Owens A, Efros AA (2020) CNN-generated images are surprisingly easy to spot... for now. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 8695–8704
193. Malik H (2019) Securing voice-driven interfaces against fake (cloned) audio attacks. In 2019 IEEE conference on multimedia information processing and retrieval (MIPR). IEEE, pp 512–517
194. Li Y, Yang X, Sun P, Qi H, Lyu S (2020) Celeb-df: a new dataset for deepfake forensics. In: IEEE Conference on Computer Vision and Patten Recognition (CVPR)
195. 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
196. Cozzolino D, Rössler A, Thies J, Nießner M, Verdoliva L (2021) ID-reveal: identity-aware DeepFake video detection. Paper presented at the international conference on computer vision, pp 15088–15097
197. Hu J, Liao X, Wang W, Qin Z (2021) Detecting compressed deepfake videos in social networks using frame-temporality two-stream convolutional network. IEEE Trans Circuits Syst Video Technol:1
198. Li X, Yu K, Ji S, Wang Y, Wu C, Xue H (2020) Fighting against deepfake: patch & pair convolutional neural networks (ppcnn). In companion proceedings of the web conference 2020, pp 88–89
199. Amerini I, Caldelli R (2020) Exploiting prediction error inconsistencies through LSTM-based classifiers to detect deepfake videos. In: Proceedings of the 2020 ACM workshop on information hiding and multimedia security, pp 97–102
200. Hosler B, Salvi D, Murray A, Antonacci F, Bestagini P, Tubaro S, Stamm MC (2021) Do Deepfakes feel emotions? A semantic approach to detecting deepfakes via emotional inconsistencies. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 1013–1022
201. Zhao T, Xu X, Xu M, Ding H, Xiong Y, Xia W (2021) Learning self-consistency for deepfake detection. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 15023–15033
202. AlBadawy EA, Lyu S, Farid H (2019) Detecting AI-synthesized speech using bispectral analysis. In: CVPR workshops, pp 104-109
203. Guo Z, Hu L, Xia M, Yang G (2021) Blind detection of glow-based facial forgery. Multimed Tools Appl 80:7687–7710. https://doi.org/10.1007/s11042-020-10098-y
204. Guo Z, Yang G, Chen J, Sun X (2020) Fake face detection via adaptive residuals extraction network. arXiv preprint arXiv:04945
205. Fu T, Xia M, Yang G (2022) Detecting GAN-generated face images via hybrid texture and sensor noise based features. Multimed Tools Appl. https://doi.org/10.1007/s11042-022-12661-1
206. Fei J, Xia Z, Yu P, Xiao F (2021) Exposing AI-generated videos with motion magnification. Multimed Tools Appl 80:30789–30802. https://doi.org/10.1007/s11042-020-09147-3
207. Singh A, Saimbhi AS, Singh N, Mittal M (2020) DeepFake video detection: a time-distributed approach. SN Comput Sci 1:212. https://doi.org/10.1007/s42979-020-00225-9
208. Han B, Han X, Zhang H, Li J, Cao X (2021) Fighting fake news: two stream network for deepfake detection via learnable SRM. IEEE Trans Biom Behav Identity Sci 3:320–331
209. Rana MS, Sung AH (2020) Deepfakestack: a deep ensemble-based learning technique for deepfake detection. In: 2020 7th IEEE international conference on cyber security and cloud computing (CSCloud)/2020 6th IEEE international conference on edge computing and scalable cloud (EdgeCom). IEEE, pp 70–75
210. Wu Z, Das RK, Yang J, Li H (2020) Light convolutional neural network with feature genuinization for detection of synthetic speech attacks. In: Interspeech 2020, 21st Annual Conference of the International Speech Communication Association. ISCA, pp 1101–1105
211. Yu C-M, Chen K-C, Chang C-T, Ti Y-W (2022) SegNet: a network for detecting deepfake facial videos. Multimedia Systems 1. https://doi.org/10.1007/s00530-021-00876-5
212. Su Y, Xia H, Liang Q, Nie W (2021) Exposing DeepFake videos using attention based convolutional LSTM network. Neural Process Lett 53:4159–4175. https://doi.org/10.1007/s11063-021-10588-6
213. Masood M, Nawaz M, Javed A, Nazir T, Mehmood A, Mahum R (2021) Classification of Deepfake videos using pre-trained convolutional neural networks. In: 2021 international conference on digital futures and transformative technologies (ICoDT2). IEEE, pp 1–6
214. Wang R, Ma L, Juefei-Xu F, Xie X, Wang J, Liu Y (2020) Fakespotter: a simple baseline for spotting ai-synthesized fake faces. In: Proceedings of the 29th international joint conference on artificial intelligence (IJCAI), pp 3444–3451
215. Pan Z, Ren Y, Zhang X (2021) Low-complexity fake face detection based on forensic similarity. Multimedia Systems 27:353–361. https://doi.org/10.1007/s00530-021-00756-y
216. Giudice O, Guarnera L, Battiato S (2021) Fighting deepfakes by detecting gan dct anomalies. J Imaging 7:128
217. Lorenzo-Trueba J, Fang F, Wang X, Echizen I, Yamagishi J, Kinnunen T (2018) Can we steal your vocal identity from the internet?: initial investigation of cloning Obama's voice using GAN, WaveNet and low-quality found data. In the speaker and language recognition workshop. ISCA, pp 240–247
218. Wang X et al (2020) ASVspoof 2019: a large-scale public database of synthetized, converted and replayed speech. Comput Speech Lang 64:101114
219. Jin Z, Mysore GJ, Diverdi S, Lu J, Finkelstein A (2017) Voco: text-based insertion and replacement in audio narration. ACM Trans Graph 36:1–13
220. Leung A NVIDIA Reveals That Part of Its CEO's Keynote Presentation Was Deepfaked. https://hypebeast.com/2021/8/nvidia-deepfake-jensen-huang-omniverse-keynote-video. Accessed August 29, 2021
221. Sotelo J, Mehri S, Kumar K, Santos JF, Kastner K, Courville A, Bengio Y (2017) Char2wav: end-to-end speech synthesis. In: 5th International Conference on Learning Representations
222. Sisman B, Yamagishi J, King S, Li H (2020) An overview of voice conversion and its challenges: from statistical modeling to deep learning. IEEE/ACM Transactions on Audio, Speech, Language Processing
223. Partila P, Tovarek J, Ilk GH, Rozhon J, Voznak M (2020) Deep learning serves voice cloning: how vulnerable are automatic speaker verification systems to spoofing trials? IEEE Commun Mag 58:100–105
224. Ping W et al (2018) Deep voice 3: 2000-speaker neural text-to-speech. Proc ICLR:214–217
225. Bińkowski M et al. (2020) High fidelity speech synthesis with adversarial networks. Paper presented at the 8th international conference on learning representations
226. Kumar K et al (2019) Melgan: generative adversarial networks for conditional waveform synthesis. Adv Neural Inf Proces Syst 32
227. Kong J, Kim J, Bae J (2020) Hifi-Gan: generative adversarial networks for efficient and high fidelity speech synthesis. Adv Neural Inf Proces Syst 33:17022–17033
228. Luong H-T, Yamagishi J (2020) NAUTILUS: a versatile voice cloning system. IEEE/ACM Trans Audio Speech Lang Process 28:2967–2981
229. Peng K, Ping W, Song Z, Zhao K (2020) Non-autoregressive neural text-to-speech. In: International conference on machine learning. PMLR, pp 7586–7598
230. Taigman Y, Wolf L, Polyak A, Nachmani E (2018) Voiceloop: voice fitting and synthesis via a phonological loop. In: 6th International Conference on Learning Representations
231. Oord A et al. (2018) Parallel wavenet: fast high-fidelity speech synthesis. In international conference on machine learning. PMLR, pp 3918–3926
232. Kim J, Kim S, Kong J, Yoon S (2020) Glow-tts: a generative flow for text-to-speech via monotonic alignment search. Adv Neural Inf Proces Syst 33:8067–8077
233. Jia Y et al. (2018) Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In: Advances in neural information processing systems, pp 4480–4490

234. Lee Y, Kim T, Lee S-Y (2018) Voice imitating text-to-speech neural networks. arXiv preprint arXiv:00927
235. Chen Y et al. (2019) Sample efficient adaptive text-to-speech. In: 7th International Conference on Learning Representations
236. Cong J, Yang S, Xie L, Yu G, Wan G (2020) Data efficient voice cloning from noisy samples with domain adversarial training. Paper presented at the 21st Annual Conference of the International Speech Communication Association, pp 811–815
237. Gibiansky A et al. (2017) Deep voice 2: multi-speaker neural text-to-speech. In: Advances in neural information processing systems, pp 2962–2970
238. Yasuda Y, Wang X, Takaki S, Yamagishi J (2019) Investigation of enhanced Tacotron text-to-speech synthesis systems with self-attention for pitch accent language. In: 2019 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 6905–6909
239. Yamamoto R, Song E, Kim J-M (2020) Parallel WaveGAN: a fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In: 2020 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 6199–6203
240. Ren Y, Ruan Y, Tan X, Qin T, Zhao S, Zhao Z, Liu T-Y (2019) Fastspeech: fast, robust and controllable text to speech. Adv Neural Inf Proces Syst 32:3165–3174
241. Toda T, Chen L-H, Saito D, Villavicencio F, Wester M, Wu Z, Yamagishi J (2016) The voice conversion challenge 2016. In: INTERSPEECH, pp 1632–1636
242. Zhao Y et al. (2020) Voice conversion challenge 2020: Intra-lingual semi-parallel and cross-lingual voice conversion. In: Proceeding joint workshop for the blizzard challenge and voice conversion challenge
243. Stylianou Y, Cappé O, Moulines E (1998) Continuous probabilistic transform for voice conversion. IEEE Trans Speech Audio Process 6:131–142
244. Toda T, Black AW, Tokuda K (2007) Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. IEEE Trans Speech Audio Process 15:2222–2235
245. Helander E, Silén H, Virtanen T, Gabbouj M (2011) Voice conversion using dynamic kernel partial least squares regression. IEEE Trans Audio Speech Lang Process 20:806–817
246. Wu Z, Virtanen T, Chng ES, Li H (2014) Exemplar-based sparse representation with residual compensation for voice conversion. IEEE/ACM Trans Audio Speech Lang Process 22:1506–1521
247. Nakashika T, Takiguchi T, Ariki Y (2014) High-order sequence modeling using speaker-dependent recurrent temporal restricted Boltzmann machines for voice conversion. In: Fifteenth annual conference of the international speech communication association
248. Ming H, Huang D-Y, Xie L, Wu J, Dong M, Li H (2016) Deep bidirectional LSTM modeling of timbre and prosody for emotional voice conversion. In: INTERSPEECH, pp 2453–2457
249. Sun L, Kang S, Li K, Meng H (2015) Voice conversion using deep bidirectional long short-term memory based recurrent neural networks. In 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 4869–4873
250. Wu J, Wu Z, Xie L (2016) On the use of i-vectors and average voice model for voice conversion without parallel data. In: 2016 Asia-Pacific signal and information processing association annual summit and conference (APSIPA). IEEE, pp 1–6
251. Liu L-J, Ling Z-H, Jiang Y, Zhou M, Dai L-R (2018) WaveNet vocoder with limited training data for voice conversion. In: INTERSPEECH, pp 1983–1987
252. Hsu P-c, Wang C-h, Liu AT, Lee H-y (2019) Towards robust neural vocoding for speech generation: a survey. arXiv preprint arXiv:02461
253. Kaneko T, Kameoka H (2018) Cyclegan-vc: Non-parallel voice conversion using cycle-consistent adversarial networks. In: 2018 26th European signal processing conference (EUSIPCO). IEEE, pp 2100–2104
254. Chou J-c, Yeh C-c, Lee H-y, Lee L-s (2018) Multi-target voice conversion without parallel data by adversarially learning disentangled audio representations. In: 19th Annual Conference of the International Speech Communication Association. ISCA, pp 501–505
255. Kaneko T, Kameoka H, Tanaka K, Hojo N (2019) Cyclegan-vc2: improved cyclegan-based non-parallel voice conversion. In: 2019 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 6820–6824
256. Fang F, Yamagishi J, Echizen I, Lorenzo-Trueba J (2018) High-quality nonparallel voice conversion based on cycle-consistent adversarial network. In 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 5279–5283
257. Hsu C-C, Hwang H-T, Wu Y-C, Tsao Y, Wang H-M (2017) Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks. Paper presented at the 18th Annual Conference of the International Speech Communication Association, pp 3364–3368
258. Kameoka H, Kaneko T, Tanaka K, Hojo N (2018) Stargan-vc: Non-parallel many-to-many voice conversion using star generative adversarial networks. In: 2018 IEEE spoken language technology workshop (SLT). IEEE, pp 266–273
259. Zhang M, Sisman B, Zhao L, Li H (2020) DeepConversion: Voice conversion with limited parallel training data. Speech Comm 122: 31–43
260. Huang W-C, Luo H, Hwang H-T, Lo C-C, Peng Y-H, Tsao Y, Wang H-M (2020) Unsupervised representation disentanglement using cross domain features and adversarial learning in variational autoencoder based voice conversion. IEEE Trans Emerg Top Comput Intell 4:468–479
261. Qian K, Jin Z, Hasegawa-Johnson M, Mysore GJ (2020) F0-consistent many-to-many non-parallel voice conversion via conditional autoencoder. In 2020 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 6284–6288
262. Chorowski J, Weiss RJ, Bengio S, van den Oord A (2019) Unsupervised speech representation learning using wavenet autoencoders. IEEE/ACM Trans Audio Speech Lang Process 27:2041–2053
263. Tanaka K, Kameoka H, Kaneko T, Hojo N (2019) AttS2S-VC: sequence-to-sequence voice conversion with attention and context preservation mechanisms. In: ICASSP 2019–2019 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 6805–6809
264. Park S-w, Kim D-y, Joe M-c (2020) Cotatron: Transcription-guided speech encoder for any-to-many voice conversion without parallel data. In: 21st Annual Conference of the International Speech Communication Association. ISCA, pp 4696–4700
265. Huang W-C, Hayashi T, Wu Y-C, Kameoka H, Toda T (2020) Voice transformer network: Sequence-to-sequence voice conversion using transformer with text-to-speech pretraining. In: 21st Annual Conference of the International Speech Communication Association. ISCA, pp 4676–4680
266. Lu H, Wu Z, Dai D, Li R, Kang S, Jia J, Meng H (2019) One-shot voice conversion with global speaker embeddings. In: INTERSPEECH, pp 669–673
267. Liu S, Zhong J, Sun L, Wu X, Liu X, Meng H (2018) Voice conversion across arbitrary speakers based on a single target-speaker utterance. In: INTERSPEECH, pp 496–500
268. Huang T-h, Lin J-h, Lee H-y (2021) How far are we from robust voice conversion: a survey. In: 2021 IEEE spoken language technology workshop (SLT). IEEE, pp 514–521

269. Li N, Tuo D, Su D, Li Z, Yu D, Tencent A (2018) Deep discriminative embeddings for duration robust speaker verification. In: INTERSPEECH, pp 2262–2266
270. Chou J-c, Yeh C-c, Lee H-y (2019) One-shot voice conversion by separating speaker and content representations with instance normalization. In: 20th Annual Conference of the International Speech Communication Association. ISCA, pp 664–668
271. Qian K, Zhang Y, Chang S, Yang X, Hasegawa-Johnson M (2019) Autovc: zero-shot voice style transfer with only autoencoder loss. In: International conference on machine learning. PMLR, pp 5210–5219
272. Rebryk Y, Beliaev S (2020) ConVoice: real-time zero-shot voice style transfer with convolutional network. arXiv preprint arXiv: 07815
273. Kominek J, Black AW (2004) The CMU Arctic speech databases. In: Fifth ISCA workshop on speech synthesis
274. Kurematsu A, Takeda K, Sagisaka Y, Katagiri S, Kuwabara H, Shikano K (1990) ATR Japanese speech database as a tool of speech recognition and synthesis. Speech Comm 9:357–363
275. Kawahara H, Masuda-Katsuse I, De Cheveigne A (1999) Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds. Speech Comm 27:187–207
276. Kamble MR, Sailor HB, Patil HA, Li H (2020) Advances in anti-spoofing: from the perspective of ASVspoof challenges. APSIPA Trans Signal Inf Process 9
277. Li X, Li N, Weng C, Liu X, Su D, Yu D, Meng H (2021) Replay and synthetic speech detection with res2net architecture. In 2021 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 6354–6358
278. Yi J, Bai Y, Tao J, Tian Z, Wang C, Wang T, Fu R (2021) Half-truth: a partially fake audio detection dataset. In: 22nd Annual Conference of the International Speech Communication Association. ISCA, pp 1654–1658
279. Das RK, Yang J, Li H (2021) Data augmentation with signal Companding for detection of logical access attacks. In: 2021 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 6349–6353
280. Ma H, Yi J, Tao J, Bai Y, Tian Z, Wang C (2021) Continual Learning for Fake Audio Detection. In: 22nd Annual Conference of the International Speech Communication Association. ISCA, pp 886–890
281. Singh AK, Singh P (2021) Detection of AI-synthesized speech using cepstral & bispectral statistics. In: 4th international conference on multimedia information processing and retrieval (MIPR). IEEE, pp 412–417
282. Gao Y, Vuong T, Elyasi M, Bharaj G, Singh R (2021) Generalized Spoofing Detection Inspired from Audio Generation Artifacts. In: 22nd Annual Conference of the International Speech Communication Association. ISCA, pp 4184–4188
283. Aravind P, Nechiyil U, Paramparambath N (2020) Audio spoofing verification using deep convolutional neural networks by transfer learning. arXiv preprint arXiv:03464
284. Monteiro J, Alam J, Falk THJCS (2020) Generalized end-to-end detection of spoofing attacks to automatic speaker recognizers. Comput Speech Lang 63:101096
285. Chen T, Kumar A, Nagarsheth P, Sivaraman G, Khoury E (2020) Generalization of audio deepfake detection. In proc. odyssey 2020 the speaker and language recognition workshop, pp 132–137
286. Huang L, Pun C-M (2020) Audio replay spoof attack detection by joint segment-based linear filter Bank feature extraction and attention-enhanced DenseNet-BiLSTM network. IEEE/ACM Trans Audio Speech Lang Process 28:1813–1825
287. Zhang Z, Yi X, Zhao X (2021) Fake speech detection using residual network with transformer encoder. In: Proceedings of the 2021 ACM workshop on information hiding and multimedia security, pp 13–22
288. Reimao R, Tzerpos V (2019) FoR: a dataset for synthetic speech detection. In international conference on speech technology and human-computer dialogue IEEE, pp 1–10
289. Zhang Y, Jiang F, Duan Z (2021) One-class learning towards synthetic voice spoofing detection. IEEE Signal Process Lett 28: 937–941
290. Gomez-Alanis A, Peinado AM, Gonzalez JA, Gomez AM (2019) A light convolutional GRU-RNN deep feature extractor for ASV spoofing detection. In: Proc Interspeech, pp 1068–1072
291. Hua G, Bengjinteoh A, Zhang H (2021) Towards end-to-end synthetic speech detection. IEEE Signal Process Lett 28:1265–1269
292. Jiang Z, Zhu H, Peng L, Ding W, Ren Y (2020) Self-supervised spoofing audio detection scheme. In: INTERSPEECH, pp 4223–4227
293. Borrelli C, Bestagini P, Antonacci F, Sarti A, Tubaro S (2021) Synthetic speech detection through short-term and long-term prediction traces. EURASIP J Inf Secur 2021:1–14
294. Malik H (2019) Fighting AI with AI: fake speech detection using deep learning. In: International Conference on Audio Forensics. AES
295. Khochare J, Joshi C, Yenarkar B, Suratkar S, Kazi F (2021) A deep learning framework for audio deepfake detection. Arab J Sci Eng 1:1–12
296. Yamagishi J et al. (2021) ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection. arXiv preprint arXiv: 00537
297. Frank J, Schönherr L (2021) WaveFake: a data set to facilitate audio deepfake detection. In: 35th annual conference on neural information processing systems
298. Dolhansky B, Bitton J, Pflaum B, Lu J, Howes R, Wang M, Ferrer CC (2020) The DeepFake detection challenge dataset. arXiv preprint arXiv:200607397
299. Jiang L, Li R, Wu W, Qian C, Loy CC (2020) Deeperforensics-1.0: a large-scale dataset for real-world face forgery detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2889–2898
300. Zi B, Chang M, Chen J, Ma X, Jiang Y-G (2020) Wilddeepfake: a challenging real-world dataset for deepfake detection. In proceedings of the 28th ACM international conference on multimedia, pp 2382–2390
301. He Y et al. (2021) Forgerynet: a versatile benchmark for comprehensive forgery analysis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 4360–4369
302. Khalid H, Tariq S, Kim M, Woo SS (2021) FakeAVCeleb: a novel audio-video multimodal deepfake dataset. In: Thirty-fifth conference on neural information processing systems
303. Ito K (2017) The LJ speech dataset. https://keithito.com/LJ-Speech-Dataset. Accessed December 22, 2020
304. The M-AILABS speech dataset. (2019). https://www.caito.de/2019/01/the-m-ailabs-speech-dataset/. Accessed Feb 25, 2021
305. Ardila R et al. (2019) Common voice: a massively-multilingual speech corpus. arXiv preprint arXiv:191206670
306. 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 preprint arXiv:180309179
307. Faceswap. https://github.com/MarekKowalski/FaceSwap/. Accessed August 14, 2020
308. Thies J, Zollhöfer M, Nießner M (2019) Deferred neural rendering: image synthesis using neural textures. ACM Trans Graph 38: 1–12
309. Abu-El-Haija S, Kothari N, Lee J, Natsev P, Toderici G, Varadarajan B, Vijayanarasimhan S (2016) Youtube-8m: a large-scale video classification benchmark. arXiv preprint arXiv: 160908675

310. Aravkin A, Burke JV, Ljung L, Lozano A, Pillonetto G (2017) Generalized Kalman smoothing: modeling and algorithms. Automatica 86:63–86
311. Reinhard E, Adhikhmin M, Gooch B, Shirley P (2001) Color transfer between images. IEEE Comput Graph 21:34–41
312. Dolhansky B, Howes R, Pflaum B, Baram N, Ferrer CC (2019) The deepfake detection challenge (dfdc) preview dataset. arXiv preprint arXiv:08854
313. Versteegh M, Thiolliere R, Schatz T, Cao XN, Anguera X, Jansen A, Dupoux E (2015) Zero resource speech challenge. In: 16th Annual Conference of the International Speech Communication Association. ISCA, pp 3169–3173
314. Mitra A, Mohanty SP, Corcoran P, Kougianos E (2021) A machine learning based approach for Deepfake detection in social media through key video frame extraction. SN Comput Sci 2:98. https://doi.org/10.1007/s42979-021-00495-x
315. Trinh L, Liu Y (2021) An examination of fairness of AI models for deepfake detection. In: Proceedings of the thirtieth international joint conference on artificial intelligence. IJCAI, pp 567–574
316. Carlini N, Farid H (2020) Evading deepfake-image detectors with white-and black-box attacks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, pp 658–659
317. Neekhara P, Dolhansky B, Bitton J, Ferrer CC (2021) Adversarial threats to deepfake detection: a practical perspective. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 923–932
318. Huang C-y, Lin YY, Lee H-y, Lee L-s (2021) Defending your voice: adversarial attack on voice conversion. In: 2021 IEEE spoken language technology workshop (SLT). IEEE, pp 552–559
319. Ding Y-Y, Zhang J-X, Liu L-J, Jiang Y, Hu Y, Ling Z-H (2020) Adversarial post-processing of voice conversion against spoofing detection. In: 2020 Asia-Pacific signal and information processing association annual summit and conference (APSIPA ASC). IEEE, pp 556–560
320. 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
321. Jung S, Keuper M (2021) Spectral distribution aware image generation. In: Proceedings of the AAAI conference on artificial intelligence, pp 1734–1742
322. Huang Y et al. (2020) FakeRetouch: evading DeepFakes detection via the guidance of deliberate noise. arXiv preprint arXiv:09213
323. Neves JC, Tolosana R, Vera-Rodriguez R, Lopes V, Proença H, Fierrez J (2020) Ganprintr: improved fakes and evaluation of the state of the art in face manipulation detection. IEEE J Sel Top Sign Process 14:1038–1048
324. Osakabe T, Tanaka M, Kinoshita Y, Kiya H (2021) CycleGAN without checkerboard artifacts for counter-forensics of fake-image detection. In: International workshop on advanced imaging technology (IWAIT) 2021. International Society for Optics and Photonics, pp 1176609
325. Huang Y et al. (2020) Fakepolisher: making deepfakes more detection-evasive by shallow reconstruction. In: Proceedings of the 28th ACM international conference on multimedia, pp 1217–1226
326. Bansal A, Ma S, Ramanan D, Sheikh Y (2018) Recycle-gan: unsupervised video retargeting. In: Proceedings of the European conference on computer vision (ECCV), pp 119-135
327. Abe M, Nakamura S, Shikano K, Kuwabara H (1990) Voice conversion through vector quantization. J Acoust Soc Jpn 11:71–76
328. Fraga-Lamas P, Fernández-Caramés TM (2020) Fake news, disinformation, and Deepfakes: leveraging distributed ledger technologies and Blockchain to combat digital deception and counterfeit reality. IT Prof 22:53–59
329. Hasan HR, Salah K (2019) Combating deepfake videos using blockchain and smart contracts. IEEE Access 7:41596–41606
330. Mao D, Zhao S, Hao Z (2022) A shared updatable method of content regulation for deepfake videos based on blockchain. Appl Intell:1–18
331. Kaddar B, Fezza SA, Hamidouche W, Akhtar Z, Hadid A (2021) HCiT: Deepfake video detection using a hybrid model of CNN features and vision transformer. In: 2021 international conference on visual communications and image processing (VCIP). IEEE, pp 1–5
332. Wodajo D, Atnafu S (2021) Deepfake video detection using convolutional vision transformer. arXiv preprint arXiv:11126
333. Wang J, Wu Z, Chen J, Jiang Y-G (2021) M2tr: Multi-modal multi-scale transformers for deepfake detection. arXiv preprint arXiv:09770
334. Deokar B, Hazarnis A (2012) Intrusion detection system using log files and reinforcement learning. Int J Comput Appl 45:28–35
335. Liu Z, Wang J, Gong S, Lu H, Tao D (2019) Deep reinforcement active learning for human-in-the-loop person re-identification. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 6122–6131
336. Wang J, Yan Y, Zhang Y, Cao G, Yang M, Ng MK (2020) Deep reinforcement active learning for medical image classification. In: International conference on medical image computing and computer-assisted intervention. Springer, pp 33–42
337. Feng M, Xu H (2017) Deep reinforecement learning based optimal defense for cyber-physical system in presence of unknown cyber-attack. In: 2017 IEEE symposium series on computational intelligence (SSCI). IEEE, pp 1–8
338. Baumann R, Malik KM, Javed A, Ball A, Kujawa B, Malik H (2021) Voice spoofing detection corpus for single and multi-order audio replays. Comput Speech Lang 65:101132
339. Gonçalves AR, Violato RP, Korshunov P, Marcel S, Simoes FO (2017) On the generalization of fused systems in voice presentation attack detection. In: 2017 international conference of the biometrics special interest group (BIOSIG). IEEE, pp 1–5

Momina Masood completed the B.Sc. (Computer Science) degree with honors and 3rd position from Fatima Jinnah Women University, Rawalpindi, Pakistan in 2015 and MS (Computer Science) degree from University of Engineering and Technology, Taxila, Pakistan in 2018. She is currently working as a Programmer at Computer Science Department, University of Engineering and Technology, Taxila, Pakistan. Her research interests include computer vision, medical image processing, multimedia forensics, and machine learning.

Marriam Nawaz has completed BSc (Software Engineering) from University of Engineering and Technology, Taxila, and secured Gold Medal. Completed her M.Sc. degree in Software Engineering from University of Engineering and Technology, Taxila, with specialization in Software Engineering Discipline.Currently enrolled in Ph.D. Program in Software Engineering at UET, Taxila. Since 2017, she has been serving as Programmer at Computer Science Department, UET Taxila. Her research interests include Image processing, Medical Image Analysis, Digital Image Forgery detection, and Deep-fakes Detection.

Khalid M. Malik (Senior Member, IEEE) is currently an Associate Professor with the School of Engineering and Computer Science, Oakland University, Rochester, MI, USA. His research interests include multimedia forensics, development of intelligent decision support systems using analysis of medical imaging and clinical text, secure multicast protocols for intelligent transportation systems, and automated ontology and knowledge graph generation. His research is supported by the National Science Foundation (NSF), Brain Aneurysm Foundation, and Oakland University. He is a founding member of the center of cybersecurity at Oakland University.

Ali Javed (M'16) received the B.Sc. degree with honors and 3rd position in Software Engineering from UET Taxila, Pakistan in 2007. He received his MS and Ph.D. degrees in Computer Engineering from UET Taxila, Pakistan in 2010 and 2016. He received Chancellor's Gold Medal in MS Computer Engineering degree. Dr. Javed is serving as an Associate Professor in Software Engineering Department at UET Taxila, Pakistan. He has served as a Postdoctoral Scholar in SMILES lab at Oakland University, MI, USA in 2019 and as a visiting PhD scholar in ISSF Lab at University of Michigan, MI, USA in 2015. His areas of interest are Multimedia Forensics, Image Processing, Computer vision, Video Content Analysis, Medical Image Processing, and Multimedia Signal Processing. He has published more than 80 papers in leading journals and conferences including the IEEE Transactions. Dr. Javed is a recipient of various research grants from HEC Pakistan, National ICT R n D Fund, NESCOM, and UET Taxila Pakistan. He has also served as an HOD in Software Engineering Department at UET Taxila in 2014. Dr. Javed got selected as an Ambassador of Asian Council of Science Editors from Pakistan in 2016. He is also a member of Pakistan Engineering Council since 2007.

Aun Irtaza has completed his PhD in 2016 from FAST-nu, Islamabad Pakistan. During his PhD he worked as a research scientist in the Gwangju Institute of Science and Technology (GIST), South Korea. He became an Associate Professor in 2017 and department of computer science chair in 2018 in the University of Engineering and Technology (UET) Taxila, Pakistan. He is currently working as visiting Associate Professor in the University of Michigan-Dearborn. His research areas include computer vision, multimedia forensics, audio-signal processing, medical image processing, and Big data analytics. He has more than 40 publications in IEEE, Springer, and Elsevier Journals.

Hafiz Malik is Professor in the Electrical and Computer Engineering (ECE) Department at University of Michigan – Dearborn. His research in the areas of automotive cybersecurity, IoT security, sensor security, multimedia forensics, steganography, steganalysis, pattern recognition, and information fusion is funded by the National Science Foundation, National Academies, Ford Motor Company, and other agencies. He has published more than 100 papers in leading journals, conferences, and workshops. He is a founding member of the Cybersecurity Center for Research, Education, and Outreach at UM-Dearborn and member leadership circle for the Dearborn Artificial Intelligence Research Center at UM-Dearborn. He is also a member of the Scientific and Industrial Advisory Board (SIAB) of the National Center of Cyber Security Pakistan. He is a member of MCity Working Group on Cybersecurity, since 2015.