

Deepfakes Generation and Detection: State-of-the-art, open challenges, countermeasures, and way forward

Momina Masood¹, Mariam Nawaz², Khalid Mahmood Malik³, Ali Javed⁴, Aun Irtaza⁵

^{1,2}Department of Computer Science, University of Engineering and Technology-Taxila, Pakistan

^{3,4}Department of Computer Science and Engineering, Oakland University, Rochester, MI, USA

⁵Electrical and Computer Engineering Department, University of Michigan-Dearborn, MI, USA

Abstract

Easy access to audio-visual content on social media, combined with the availability of modern tools such as Tensorflow or Keras, open-source trained models, and economical computing infrastructure, and the rapid evolution of deep-learning (DL) methods, especially Generative Adversarial Networks (GAN), have made it possible to generate deepfakes to disseminate disinformation, revenge porn, financial frauds, hoaxes, and to disrupt government functioning. The existing surveys have mainly focused on deepfake video detection only. No attempt has been made to review approaches for detection and generation of both audio and video deepfakes. 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 for the detection and generation of both audio and video deepfakes. For each category of deepfake, we discuss information related to manipulation approaches, current public datasets, and key standards for the performance evaluation of deepfake detection techniques along with their results. Additionally, we also discuss open challenges and enumerate future directions to guide future researchers on issues that need to be considered to improve the domains of both the deepfake generation and detection. This work is expected to assist the readers in understanding the creation and detection mechanisms of deepfake, along with their current limitations and future direction.

Keywords Artificial intelligence, Deepfakes, Deep learning, Face swap, Lip synching, Puppet master, Speech synthesis, Voice replay.

1 Introduction

The availability of economical digital smart devices like cellphones, tablets, laptops, and digital cameras has resulted in the exponential growth of multimedia content (e.g. images and videos) in cyberspace. Additionally, the evolution of social media over the last decade has allowed people to share captured multimedia content rapidly, leading to a significant increase in multimedia content generation and ease of access to it. At the same time, we have witnessed tremendous advancement in the field of ML with the introduction of sophisticated algorithms that can easily manipulate multimedia content to spread disinformation online through social media platforms. Given the ease with which false information may be created and spread, it has become increasingly difficult to know the truth and trust the information, which may result in harmful consequences. Moreover, today we live in a “post-truth” era, where a piece of information or disinformation is utilized by malevolent actors to manipulate public opinion. Disinformation is an active measure that has the potential to cause severe damage: election manipulation, creation of warmongering situations, defaming any person, etc. In recent times, deepfake generation has significantly advanced and it could be used to propagate disinformation around the globe, and may pose a severe threat, in the form of fake news, in the future. Deepfakes are synthesized, AI-generated, videos and audio. The use of videos as evidence in every sector of litigation and criminal justice proceedings is currently the norm. A video admitted as evidence must be authentic and its integrity must be verified. This is expected to become a challenging task, especially as deepfake generation becomes more sophisticated. Once the deepfakes have been created, the further use of powerful, sophisticated, and easy-to-use manipulation tools (e.g. Zao[1], REFACE[2], FaceApp[3], Audacity [4], Soundforge [5]) could make authentication and integrity verification of generated videos an even more difficult task.

Deepfakes videos 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 deepfakes. In face-swap deepfakes, the face of the source person is replaced with the target person to generate a fake video of the target person, trying to portray actions to the target person which in reality the source person has done. Face-swap-oriented deepfakes are usually generated to target the popularity or reputation of famous personalities by showing them in scenarios in which they never appeared [6], and to damage reputations in the face of the public, for example, in non-consensual pornography [7]. In lip-synching

based deepfakes, the movements of the target person's lips are transformed to make them consistent with some specific audio recording. Lip-syncing is generated with the aim of showing an individual speaking in a way in which the attacker devises the victim to speak. With puppet-master, deepfakes are created by mimicking 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 the full-body, [8] in a video, and to animate according to the impersonator's desire. Face synthesis and attribute manipulation involves the generation of photo-realistic face images and facial attribute editing. This manipulation is generated to spread disinformation on social media using fake profiles. Lastly, audio deepfakes, also known as voice cloning, focus on the generation of the target speaker's voice using deep learning techniques to portray the speaker saying something they have not said [9, 10].

Unlike deepfake videos, less attention has been paid to the detection of audio deepfakes. In the last few years voice cloning has also become very sophisticated. Voice cloning is not only a threat to automated speaker verification systems, but also to voice-controlled systems deployed in Internet of Things (IoT) settings [9, 10]. Voice cloning has tremendous potential to destroy public trust and to empower criminals to manipulate business dealings or private phone calls. For example, recently three cases were reported in which bank robbers used voice cloning of a company executive's speech to dupe their subordinates into transferring hundreds of thousands of dollars into a secret account [11]. The integration of voice cloning into deepfakes is expected to become a unique challenge for deepfake detection. Therefore, it is important that, unlike current approaches that focus only on detecting video signal manipulations, audio forgeries should also be examined.

There are no existing recently published surveys on deepfake generation and detection that focus on the generation and detection of both the audio and video modalities of deepfakes. Most of the existing surveys focus only on reviewing deepfakes images, and video detection. In [12], the main focus was on generic image manipulation and multimedia forensic techniques. However, this work has not discussed deepfake generation techniques. In [13], an overview of face manipulation and detection techniques was presented. Another survey [14] was presented on visual deepfakes detection approaches, but does not discuss audio cloning and its detection. The latest work presented by Mirsky et al. [15] gives an in-depth analysis of visual deepfake creation techniques, however, deepfake detection approaches are only briefly discussed. Moreover, this work [15] lacks a discussion of audio deepfakes. According to the best of our knowledge, this paper is the first attempt to provide a detailed analysis and review of both the audio and visual deepfake detection techniques, as well as generative approaches. The following are the main contributions of our work:

- i. To give the research community an insight into 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.

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 of deepfakes is briefly discussed. Section 4 presents the audio-visual deepfake generation chain, and deepfake detection techniques are discussed in Section 5. Section 6 presents the available datasets used for both audio and video deepfakes detection. In Section 7, we discuss the possible future trends of both deepfakes 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 by influential society representatives to fabricate original information to achieve the planned political or financial objectives. It is expected to become the main process of intentionally spreading manipulated news to affect public opinion or obscure reality. Because of the extensive use of social media platforms it is now very easy to spread false news [16]. Although all categories of fake multimedia (i.e. fake news, fake images, and fake audio) could be sources of disinformation and misinformation, audiovisual-based deepfakes are expected to be much more devastating. Historically, deepfakes were created to make famous personalities controversial among their fans. For example, in 2017 a celebrity faced such a situation when a fake pornographic video was circulated in cyberspace. This is evidence that deepfakes can be used to damage the reputations, i.e. character's assassination of renowned people to defame them [14], blackmailing individuals for monetary benefits, or to create political or religious unrest by targeting politicians or religious scholars with fake videos/speeches [17], etc. This damage is not limited to targeting individuals; rather deepfakes can be used to

manipulate elections, create warmongering situations by showing fake videos of missiles launched to destroy the enemy state, or used to deceive military analysts by portraying fake information, like showing a fake bridge across the river, to mislead troop deployment, and so on.

The deepfakes are expected to advance the following current sources of disinformation and misinformation to the next level.

Trolls: Independent Trolls are hobbyists who spread inflammatory information to cause disorder and reactions in society by playing with the emotions of people [18]. For example, posting audio-visual manipulated racist or sexist content and infuriating individuals may promote hatred among the individuals. Similarly, during the 2020 election campaign of US President Donald Trump, conflicting narratives about Trump and Biden were circulated on social media, contributing to an environment of fear [19]. Opposed to independent trolls who spread false information for their own satisfaction, hired trolls will perform the same job for monetary benefits. Different actors, like political parties, businessmen, and companies routinely hire people to forge news related to their competitors and spread it in the market [20]. For example, according to a report published by Western intelligence [21], Russia is running “troll farms,” where trolls are trained to affect conversations related to national or international issues. According to these reports, deepfake videos generated by hired trolls are the newest weapon in the ongoing fabricated news war that can bring a more devastating effect on society.

Bots: Bots are automated software or algorithms used to spread fabricated or misleading content among the people. A study published in [22, 23] concludes that during the US election campaign-2016, bots were employed to generate one-fifth of the tweets during the last month of the campaign. The emergence of deepfakes has empowered the negative impact of bots i.e. recently, a messaging app named telegram [24] used bots to produce nude pictures of women, which is under investigation by Italian authorities.

Conspiracy Theorists: Conspiracy Theorists can 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 [25]. They believe that certain prominent communities are running the public while concealing their activities, like conspiracy theories about a Jewish plan to control the world [25, 26]. Moreover, recently, several conspiracy theorists have connected the current COVID pandemic with the USA and China. In such a situation, 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. Because of the extensive usage of social media, the Hyper-partisan media is one of the biggest potential incubator for spreading fabricated news among the people [27]. The convincing AI-generated fake content can assist these bloggers to easily spread disinformation to attract visitors or increase views. As social platforms are largely independent and ad-driven mediums, spreading fabricated information may purely be a profit-making strategy.

Politicians: One of the main sources of disinformation is the political parties themselves, which may spread manipulated information for point scoring. Due to the large number of followers on social platforms, politicians are central nodes in online networks. So, politicians can use their fame and public support to spread false news among their followers. To defame opponent parties, politicians can use deepfakes to post controversial content about their competitors on conventional media [25].

Foreign Governments: As the Internet has converted the world into a “Global Village,” it is easy for conflicting countries to spread false news to advance their agendas abroad. Their motive is to target the reputation of a country in the rest of the world. Many countries are running government-sponsored social media accounts, websites, and applications, contributing to political propaganda globally. Particularly, the governments of China, Israel, Turkey, Russia, UK, Ukraine, and North Korea etc. are believed to be involved in using ‘digital footsoldiers’ to smear opponents, spreading disinformation and posting fake texts for ‘pocket money’. These countries run numerous official social sites over various online platforms like Twitter, Instagram, and Facebook, etc. [28]. The ability to doctor multimedia content has become so easy that private actors maybe able to initiate foreign attacks on their own to increase the stress among countries.

3 DeepFake 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 Abraham Lincoln [29]. Usually, such manipulation is accomplished by adding (splicing), removing (inpainting), and replicating (copy-move) the objects within or between two images [12]. Then, suitable post-processing steps like scaling, rotating, and color adjustment are applied to improve visual appearance, scale, and perspective coherence.

Aside from these traditional manipulation methods, advancements in Computer Graphics and DL techniques now offer a variety of different automated approaches for digital manipulation with better semantic consistency. The recent

trend involves the synthesis of videos from scratch using autoencoders, or GAN, for different applications [30] and, more specifically, photorealistic human face generation based on any attribute [31-34]. 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 [35]. 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. Fig. 1 shows the evolution of deepfakes over the years. An early notable academic project was Video Rewrite Program [36], intended for applications in movie dubbing, published in 1997. It was the first software used 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 [14]. Another notorious deepfake case was the release of the deepNude application that allowed users to generate fake nude images [37]. This was the beginning of when deepfakes gained wider recognition within a large community. Today deepfake technology/applications, i.e. FakeApp [38], FaceSwap [39], and ZAO [1] 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 [40] 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 1. Contemporary academic projects that lead to the development of deepfake technology are Face2Face [33] and Synthesizing Obama [32], published in 2016 and 2017 respectively. Face2Face [33] 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 source facial expressions. Synthesizing Obama [32] is a video rewrite 2.0 program, used to modify the mouth movement in the video footage of a person to depict the person saying the words contained in an arbitrary audio clip. These works [32, 33] are focused on the manipulation of the head and facial region only. Recent development expands the application of deepfakes to the entire body, [8, 41] and the generation of deepfakes from a single image [42-44].

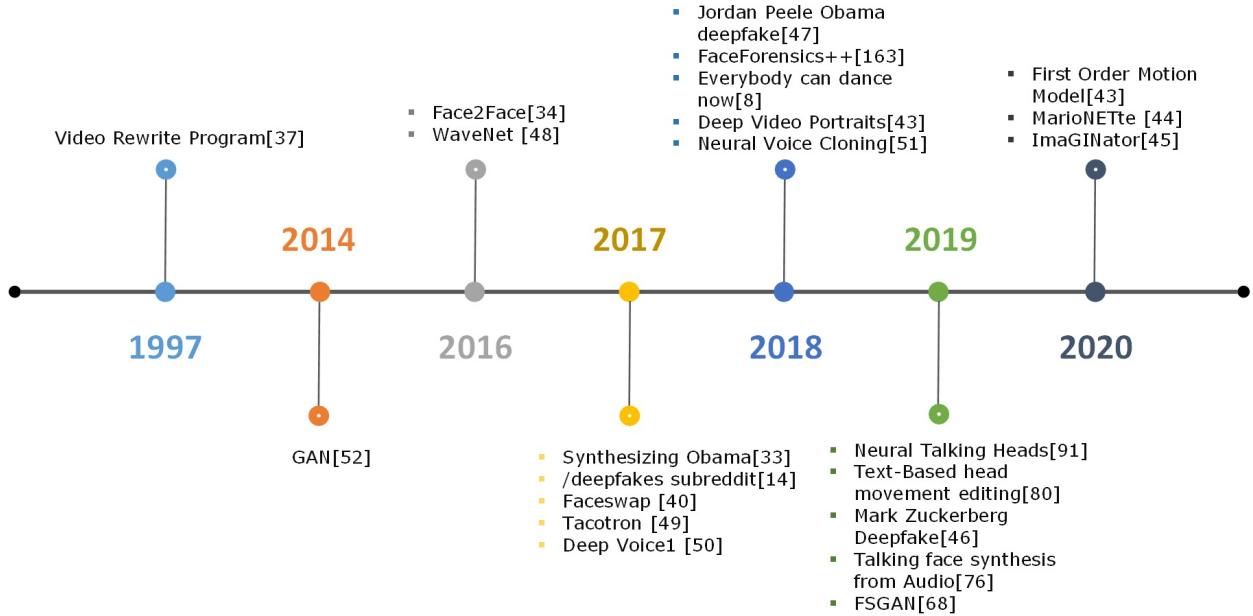


Figure 1: The timeline of Deepfakes evolution

Most deepfakes currently present on social platforms like YouTube, Facebook or Twitter may be regarded as harmless, entertaining, or artistic. However there are also some examples where deepfakes have been used for revenge porn, hoaxes, for political or non-political influence, and financial fraud [35, 45]. In 2018, a deepfake video went viral online in which former U.S. President Barak Obama appeared to insult the current president, Donald Trump [46]. In June 2019, a fake video of Facebook CEO Mark Zuckerberg was posted to Instagram by the Israeli advertising company “Canny” [45].

Table 1: 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-synching			
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	Neocortex, Inc	Proprietary
Reflect	Mobile app	Neocortex, 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 pairs of encoder-decoder. Shared encoder parameters.
DFaker	Open source implementation	github.com/dfaker/df	For face reconstruction DSSIM loss function [34] is utilized. Keras library-based implementation.
DeepFaceLab	Open source implementation	github.com/iperov/DeepFaceLab	- provide several face extraction methods, e.g. dlib, MTCNN, S3FD etc. - Extend different Faceswap model i.e. H64, H128, LIAEF128, SAE [33].
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, used 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
Imitator	Mobile app		Proprietary (AI based)
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 Cloning			
Overdub	Commercial Web App	www.descript.com/overdub	Proprietary (AI based)
Respeecher	Commercial Web App	www.respeecher.com/	Combined 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)

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 [47], Tacotron [48], and deep voice1 [49]. Fake audio-assisted financial scams have increased significantly in 2019 due to 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 [11]. 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 [50].

4 Deepfake Generation

This section provides an in-depth analysis of existing state-of-the-art methods for audio and visual deepfake generation. A review for each category of deepfake creation 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 deepfake creation.

AI-generated synthetic media has become pervasive in our digital society. Mainly, deep learning architectures, such as GANs [51] or Variational Autoencoder (VAEs) [52], are used to create the multimedia content that includes hyper-realistic images, videos, and even audio [51]. These models have a variety of applications in the real-world, such as generation of text-to-speech [53], text-to-image [54], and training data for medical imaging [55].

Creation of deepfakes mainly falls into the following categories. (i) face swap, (ii) lip-syncing, (iii) puppet-mastery, iv) face synthesis and attribute manipulation, and v) audio deepfakes. In face-swap [56], or face replacement, the face of the person in the source video is automatically replaced by the face in the target video, as shown in Fig. 2(a). In lip-syncing, the source video is modified such that it generates a video with a consistent mouth region using an arbitrary audio recording [32]. Puppet-master, also known as face reenactment [57], is a technique in which the facial expression and movements of the person in the target video or image are controlled by the person in the source video. In puppet-master deepfakes, a performer sitting in front of a camera guides the motion and deformation of a face appearing in a video or image, as shown in Fig. 2(b). This face swap approach is based on replacing the source identity with the target identity (identity manipulation), whereas the puppet-mastery and lip-syncing approaches deal with the manipulation of facial expressions. Face synthesis and attribute manipulation focuses on the creation of fake facial images, and attribute fabrication. Table 2 presents an overview of different state-of-the-art visual deepfake creation methods. Audio deepfakes deal with the modification and creation of the target’s speech. Modern advancements may lead to the manipulation of both audio and visual content [38, 58] or the movement of the entire body within a video clip [59].

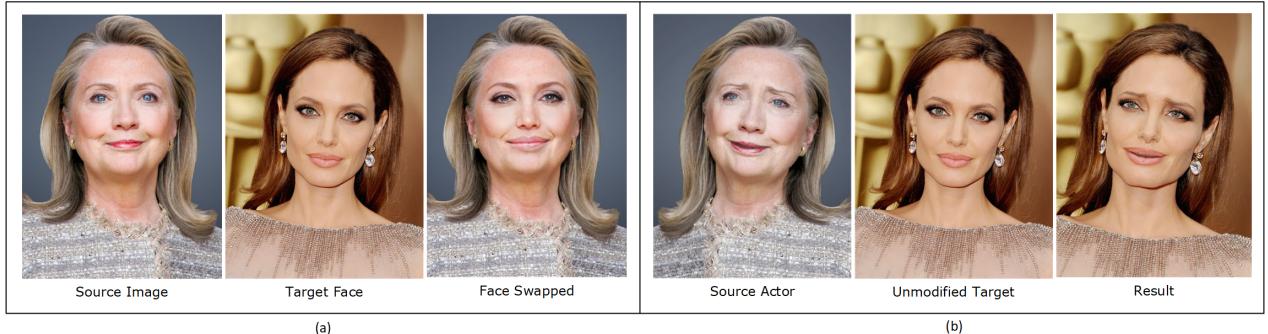


Figure 2: (a) Face Swapping (b) Facial Reenactment

4.1 Face-swap

Visual manipulation is nothing new; images and videos have been forged since the beginning. Traditional face-swap approaches [60-62] generally take three steps to perform a face-swap operation. First, these tools detect the face in source images and then select a candidate face image from the facial library that is similar to input facial appearance and poses. 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 ranks the blended candidate replacement by computing a match distance over the overlap region. These approaches may offer good results under certain conditions 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 e.g. 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. At the same time, deepfakes showed how these approaches can be applied 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 [14]. 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 [38] and FaceSwap [39] has 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. Usually, an encoder is used to extract the latent features of a face from the image and then 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 decoder of the target image is 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. 3 is an example of a deepfake creation where 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 [1], REFACE [2], and FakeApp [38] 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 as less tech-savvy users can 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 [39], DFaker [63], DeepFaceLab [40], DeepFake-tf [64], and FaceSwapGAN [65], leading to the creation of a growing number of synthesized media clips.

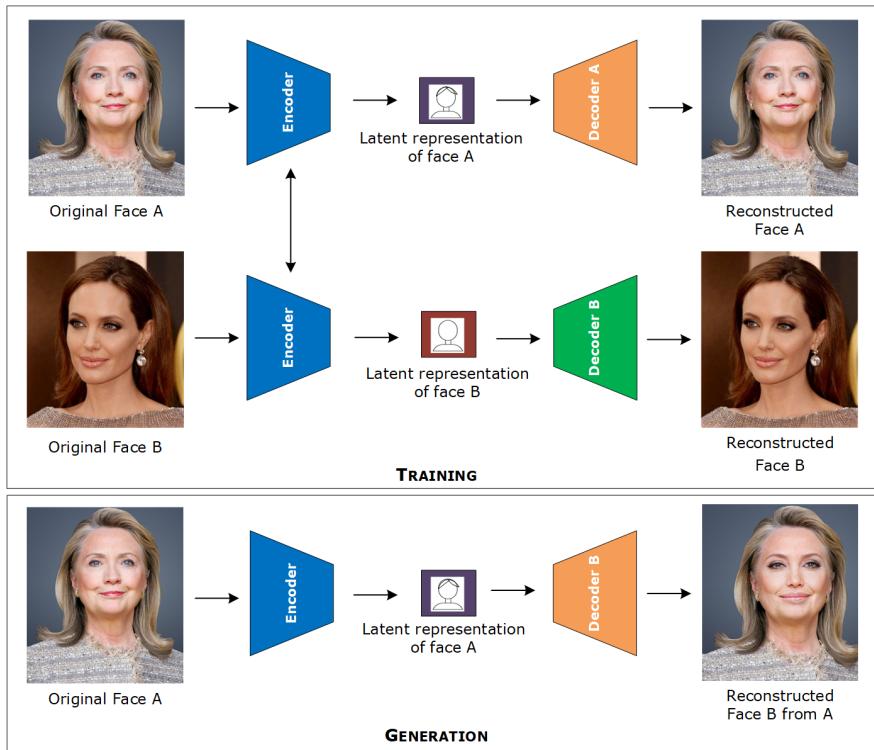


Figure 3: Creation of a Deepfake using an auto-encoder and decoder. The same encoder-decoder pair is used to learn the latent features of the faces during training, while during generation decoders are swapped, such that latent face A is subjected to decoder B to generate face A with the features of face B

Until recently, most of the research focused on advances in face-swapping technology, either using a reconstructed 3D morphable model (3DMM) [56, 66], or GANs based model [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 that style 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 [60], 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 [56] presented a method that used a full convolution network (FCN) for face segmentation and replacement while a 3DMM was established to estimate

facial geometry and corresponding texture. Then the face reconstruction was performed on a target image by adjusting the model parameters. These approaches [56, 66] have the limitation of subject-specific or pair-specific training. Recently subject agnostic approaches have been proposed to address this limitation.

In [65], an improved deepfake using GAN was proposed which adds adversarial loss and perceptual loss to VGGface implemented in the auto-encoder architecture [39]. The addition of VGGFace perceptual loss made the direction of the eyeball 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 [38, 39] as they work without being explicitly trained on subject-specific images. Moreover, the iterative nature makes them well-suited for face manipulation tasks such as generating realistic images of fake faces.

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 was later combined to generate the swapped face. However, these approaches [68, 69] hardly preserves target attributes like target occlusion and illumination conditions.

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 last layer of a facial recognition classifier was used to encode the source identity and the target attributes, with feature maps being obtained via the U-Net decoder. These encoded features were passed to a novel generator with cascaded Adaptive Attentional Denormalization layers inside residual blocks which adaptively adjusted the identity region and target attributes. Finally, another network was used to fix occlusion inconsistencies and refine the results.

4.2 Lip syncing

The Lip-syncing approach involves synthesizing a video of a target identity such that the mouth region in the manipulated video is consistent with arbitrary audio input (Fig. 4). A key aspect of synthesizing a visual speech 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 hearing-impaired persons understand a scenario by lip-reading from a video created using the genuine audio.

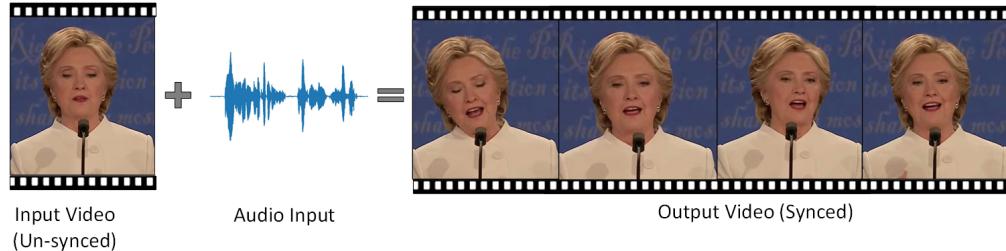


Figure 4: A visual representation of lip-syncing of an existing video to an arbitrary audio clip

Existing works on lip-syncing [71, 72] require the reselection of frames from a video or transcription, along with target emotions, to synthesize the lip's motion. These approaches are limited to a dedicated emotional state or don't generalize well to unseen faces. However, the DL models are capable of learning and predicting the movements from audio features. Suwajanakorn et al. [32] proposed an approach to generate a photo-realistic lip-synced video using a target's video and an arbitrary audio clip as input. The recurrent neural network (RNN) based model was employed to learn 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 was performed on the lower facial regions i.e. mouth, chin, nose, and cheeks. This approach applied 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 was considered as a case study due to the sufficient availability of online video

footage. Thus, this model is required to retrain for each individual. The Speech2Vid [73] model took an audio clip and a static image of a target subject as input, and generated a video that is lip-synced with the audio clip. This model used the Mel Frequency Cepstral Coefficients (MFCC) features extracted from the audio input and fed them into a CNN based encoder-decoder. As a post-processing step, a separate CNN was used for frame deblurring and sharpening 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 emotional facial expressions.

Vougioukas et al. [74] used a temporal GAN, consisting of an RNN, to generate a photorealistic video directly from a still image and speech signal. The resulting video included synchronized lip movements, eye-blinking, and natural facial expression without relying on manually handcrafted audio-visual features. Multiple discriminators were 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 [75], an adversarial learning method was employed to learn the disentangled audio-visual representation. The speech encoder was trained to project both the audio and visual representations into the same latent space. The advantage of using a disentangled representation was that both the audio and video could serve as a source of speech information during the generation process. As a result, it was possible to generate realistic talking face sequences on an arbitrary identity with synchronized lip movement. Garrido et al. [76] presented a Vdub system that captures the high-quality 3D facial model of both the source and the target actor. The computed facial model was used to photo-realistically reconstruct a 3D mouth model of the dubber to be applied on the target actor. An audio channel analysis was 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 [77] a face-to-face translation method, LipGAN, was proposed to 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. [78] proposed a wav2lip speaker-independent model that can accurately synchronize the lips movement in a video recording with 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.

The recent approaches can synthesize photo-realistic fake videos from speech (speech-to-video) or text (text-to-video) with convincing video results. The methods proposed in [32, 79] can edit 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 only. In [80] a VAE based framework was proposed to synthesize full pose video with facial expressions, gestures, and body posture movements from given audio.

4.3 Puppet-master

Puppet-master, also known as face reenactment, is another common variation of deepfakes that manipulates the facial expressions of a person e.g., transferring the facial gestures, eye, and head movements to an output video which reflect those of the source actor. Puppet-mastery aims to deform the person's mouth movement to make fabricated content. Facial reenactment has various applications, i.e. 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. [81, 82] presented the first real-time facial expressions transfer method from an actor to a target person. A commodity RGB-D sensor was used to track and reconstruct the 3D model of a source and target actor. 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 [33] is an advanced form of facial reenactment technique as presented in [81]. This method works in real-time and is 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 creates a convincing and instantaneous re-rendering of the 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 [83] using an inertial measurement unit [84].

GANs have been successfully applied for facial reenactment due to their ability to generate photo-realistic images. Pix2pixHD [85] produces high-resolution images with better fidelity by combining multi-scale conditional GANs (cGAN) architecture using a perceptual loss. Kim et al. [86] proposed an approach that allows the full 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 to predict the synthetic rendering into photo-realistic video frames. This approach requires training the videos for target identity. Wu et al. [87] proposed ReenactGAN which encodes the 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 [88] 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 [89] used facial landmarks along with the self-attention mechanism for facial reenactment. This approach introduced a triple consistency loss to minimize visual artifacts but requires the images to be synthesized with a frontal facial view for further processing. These models [89-90] require a large amount of training data for target identity to perform well at oblique angles or they will lack 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 [34], 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 target expression, was proposed. X2face used 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 frame to the vector map. The driving network was crafted to interpolate face representation from the embedded network to produce target expressions. Zakharov et al. [90] presented 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 was obtained by averaging the target's expressions and associated landmarks from different frames. Then a pix2pixHD [85] GAN was used to generate the target identity using source landmarks as input, and identity encoding via AdaIN layers. This approach works well at oblique angles and directly transfers the expression without requiring intermediate boundary latent space or interpolation map, as in [34]. Zhang et al. [91] proposed an auto-encoder-based structure to learn the latent representation of the target's facial appearance and source's face shape. These features were used as input to SPADE residual blocks for the face reenactment task, which preserved the spatial information and concatenated 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 FaRGAN [92], learnable features from convolution layers were used as input to the SPADE module instead of using multi-scale landmark masks, as in [91]. 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 [43] was proposed to mitigate identity leakage by employing attention block and target feature alignment. This helped the model to accommodate the variations between face structures better. Finally, the identity was retained by using a novel landmark transformer, influenced by the 3DMM facial model [93]. [90, 92]

FSGAN [67] can perform both the facial replacement and reenactment with occlusion handling. For reenactment, a pix2pixHD [85] 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 was trained recursively where output was iterated multiple times for incremental interpolation from source to target landmarks. The results were 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.

Thies et al. [84] extended the facial expression reenactment concept to drive the movement of the torso, eye, and head of the target by using parametric models of these three features. In the next few years, photo-realistic full-body reenactment [8] videos will also be viable, where the target's expression, along with mannerism, 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 the fabricated content completely [94]. These progressions enable the real-time manipulation of facial expressions and motion in videos, while making it challenging to distinguish between real and synthesized video.

4.4 Face Synthesis and Attribute Editing

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. Face manipulation can be broadly grouped into two categories: face generation and face attribute editing. Face generation involves the synthesis of photorealistic images of a human face that doesn't exist in real life. In contrast, face attribute editing involves altering the facial appearance of an existing sample by modifying the 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 have been using commercially available AI-based face editing and mobile applications such as FaceApp [3] to automatically alter the appearance of an input image.

The tremendous evolution in deep generative models has made them widely adopted tools for image synthesis and editing. Generative deep learning models, i.e. GAN [51] and VAE [95], have been successfully used to generate photorealistic 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, like the synthesis of photorealistic fake image for social network accounts with a false digital identity to spread misinformation. Several approaches have been proposed to generate realistic-looking facial images that humans are unable to recognize as to whether they are real or synthesized. Fig. 5 shows synthetic facial images and the improvement in their quality between 2014 and 2019, that are nearly indistinguishable from real photographs.



Figure 5: Increasingly improving improvements in the quality of synthetic faces, as generated by variations on GANs. In order, the images are from papers by Goodfellow et al. (2014) [51], Radford et al. (2015) [96], Liu et al. (2016) [97], Karras et al. (2017) [98], and Style-based (2018 [99], 2019 [100])

Since the emergence of GAN [51] in 2014, significant efforts have been made to improve the quality of synthesized images. The images generated using the first GAN model [51] were low-resolution and not very convincing. DCGAN [96] 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. [97] 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. [98] presented ProGAN, a training methodology for GANs, that employed an adaptive mini-batch size which progressively increased the resolution, depending on the current output resolution, by adding layers to the networks during the training process. StyleGAN [99] is an improved version of ProGAN [98]. 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 is that the intermediate latent vector is free from any certain distribution restriction, and this reduces the correlation between features (disentanglement). The layers of the generator network are controlled via an AdaIN operation which helps decide the features in the output layer. Compared to [51, 96, 97], StyleGAN [99] achieved state-of-the-art high resolution in the generated images i.e., 1024×1024 , with fine details. StyleGAN2 [100] 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. [101] presented a Two-Pathway Generative Adversarial Network (TP-GAN) that could simultaneously perceive global structures and local details,

like humans, and synthesize a high-resolution frontal view facial image from a single ill-posed face image. Image synthesis using this approach preserves the identity under large pose variations and illumination. Zhang et al. [102] 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 [103], authors proposed BigGAN architecture, which uses 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 is 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. [104] proposed a stacked GAN (StackGAN) model to generate high-resolution images (e.g., 256×256) with details based on a given textual description.

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 expression, etc., of the 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. 6. Perarnau et al. [105] introduced the Invertible Conditional GAN (IcGAN), which uses an encoder in combination with cGANs for face attribute editing. The encoder maps the input face image into latent representation and attributes manipulation vector and cGAN reconstructs the 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 [106], a Fader Network was presented, where an encoder-decoder architecture was trained in an end-to-end manner which generated 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 original fine details in the generated image.

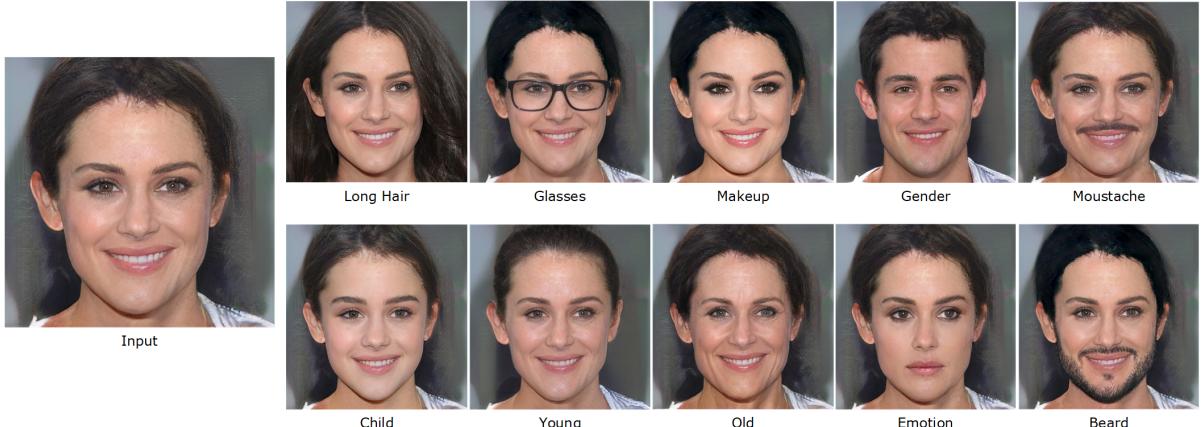


Figure 6: Examples of different face manipulations: original sample (Input) and manipulated samples

Prior studies [105, 106] have been focused on handling image-to-image translations between two domains. These methods required the different generator to be trained independently to handle translations between each pair of image domains and thus limits their practical usage. StarGAN [31], an enhanced approach, is 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. However, this approach adds some undesired visible artifacts in the facial skin such as the uneven color tone in the output image. The recently proposed StarGAN-v2 [107] achieved state-of-the-art visual quality of the generated images as compared to [31] by adding a random Gaussian noise vector into the generator. In AttGAN [108], an encoder-decoder architecture was proposed that considers the relationship between attributes and latent representation. Instead of imposing an attribute independent constraint on latent representation like in [105, 106], an attribute classification constraint was applied to the generated image 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, adds unwanted changes and blurriness, and generates low quality edited results. Liu et al. [109] 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 focuses on the attribute-specific region and does not guarantee good preservation of the details in attribute-irrelevant regions.

Table 2: An overview of visual Deepfake generation techniques

Reference	Technique	Features	Dataset	Output Quality	Limitations
FaceSwap					
Faceswap [39]	Encoder-decoder	Facial landmarks	Private	256×256	<ul style="list-style-type: none"> Blurry results due to lossy compression Lack of pose, facial expression, gaze direction, hairstyle, and lighting Requires massive no. of target images
FaceSwapGAN [65]	GAN	VGGFace	VGGFace	256×256	<ul style="list-style-type: none"> Lack of texture details and generate overly smooth results
DeepFaceLab [110]	Encoder-decoder	Facial landmarks	Private	256×256	<ul style="list-style-type: none"> Fails to blend very different facial hues Requires target training data
Fast Face-swap [66]	CNN	VGGFace	<ul style="list-style-type: none"> CelebA (200,000 images) Yale Face Database B (different pose and lighting conditions) 	256×256	<ul style="list-style-type: none"> Works for a single person only Gives better result for frontal face view Lack of skin texture details, e.g., smooth results and Facial Expression transfer Lack of occluding objects i.e. glasses
Nirkin et al. [56]	FCN-8s-VGG architecture	<ul style="list-style-type: none"> Basel Face Model to represent faces 3DDFA model for expression 	IARPA Janus CS2 (1275 face videos)	256×256	<ul style="list-style-type: none"> Poor results in case of different image resolutions Fails to blend very different facial hues
Chen et al. [111]	VGG-16 net	68 facial landmarks	Helen (2330 images)	256×256	<ul style="list-style-type: none"> Provide more realistic results but sensitive to variation in posture and gaze
FSNet [69]	GAN	Facial landmarks	CelebA	128×128	<ul style="list-style-type: none"> Sensitive to variation in angle
RSGAN [68]	GAN	Facial landmarks, segmentation mask	CelebA	128×128	<ul style="list-style-type: none"> Sensitive to variation in angle, occlusion, lightning Limited output resolution
FaceShifter [70]	GAN	Attributes (face, occlusions, lighting or styles)	<ul style="list-style-type: none"> VGG Face CelebA-HQ FFHQ 	256×256	<ul style="list-style-type: none"> Stripped artifacts
Lip-syncing					
Suwajanakorn et al. [32]	RNN (single-layer unidirectional LSTM)	<ul style="list-style-type: none"> Mouth landmarks (36-D features) MFCC audio features (28-D) 	Youtube videos (17 hours)	2048×1024	<ul style="list-style-type: none"> Requires large amount of training data for target person. require retraining for each identity. Sensitive to the 3D movement of head No direct control over facial expressions
Speech2Vid[73]	Encoder-decoder CNN	<ul style="list-style-type: none"> VGG-M network MFCC audio features 	<ul style="list-style-type: none"> VGG Face LRS2 (41.3-hour video) VoxCeleb2 (test) 	109×109	<ul style="list-style-type: none"> lacks the synthesis of emotional facial expressions
Vougioukas et al. [74]	Temporal GAN	MFCC audio features	<ul style="list-style-type: none"> GRID TCD TIMIT 	96×128	<ul style="list-style-type: none"> lacks the synthesis of emotional facial expressions flickering and jitter sensitive to large facial motions
Zhou et al. [75]	Temporal GAN	Deep audio-video features	<ul style="list-style-type: none"> LRW MS-Celeb-1M 	256×256	<ul style="list-style-type: none"> lacks the synthesis of emotional facial expressions
Vdub [76]	3DMM	<ul style="list-style-type: none"> 66 facial feature points MFCC features 	Private	1024×1024	<ul style="list-style-type: none"> Requires video of the target
LipGAN [77]	GAN	<ul style="list-style-type: none"> VGG-M network MFCC features 	LRS 2	1280×720	<ul style="list-style-type: none"> visual artifacts and temporal inconsistency unable to preserve source lip region characteristics
Wav2Lip[78]	GAN	<ul style="list-style-type: none"> Mel-spectrogram representation 	LRS2	1280×720	<ul style="list-style-type: none"> lacks the synthesis of emotional facial expressions
Face reenactment					
Face2Face [33]	3DMM	<ul style="list-style-type: none"> parametric model Facial landmark features 	customized	1024×1024	<ul style="list-style-type: none"> Sensitive to facial occlusions
Kim et al. [86]	cGAN	parametric model of the face (261 parameters/frame)	customized	1024×1024	<ul style="list-style-type: none"> 1-3 min. video of target Sensitive to facial occlusions

ReenactGAN [87]	GAN	Facial landmark features	<ul style="list-style-type: none"> ▪ CelebV dataset ▪ WFLW Dataset ▪ Helen, DISFA 	256×256	<ul style="list-style-type: none"> ▪ 30 min. video of target ▪ Lack of gaze adaption
GANimation [88]	GAN (2 Encoder- 2 Decoder)	AUs	<ul style="list-style-type: none"> ▪ EmotioNet dataset ▪ RaFD dataset 	128×128	<ul style="list-style-type: none"> ▪ Lack of pose and gaze adaption
GANnotation [89]	GAN	Facial landmark features	<ul style="list-style-type: none"> ▪ 300-VWChallenge dataset ▪ BP4D dataset ▪ Helen, LFPW, AFW, IBUG, and a subset of multiple datasets 	128x128	<ul style="list-style-type: none"> ▪ Lack of gaze adaption
X2face [34]	2Encoder-2Decoder	<ul style="list-style-type: none"> ▪ Facial landmark features ▪ 256-D audio features 	<ul style="list-style-type: none"> ▪ VGG Face dataset ▪ VoxCeleb dataset ▪ AFLW dataset 	256×256	<ul style="list-style-type: none"> ▪ Wrinkle artifacts ▪ Lack of gaze adaption
Zakharov et al. [90]	GAN (1Encoder-2Decoder)	Facial landmark features	VoxCeleb dataset	256×256	<ul style="list-style-type: none"> ▪ Sensitive to source identity leakage ▪ Lack of gaze adaption
Zhang et al. [91]	GAN (1Encoder-2Decoder)	Appearance and shape feature Map	<ul style="list-style-type: none"> ▪ VGG Face dataset ▪ WFLW ▪ EOTT dataset ▪ CelebA-HQ dataset ▪ LRW dataset. 	256×256	<ul style="list-style-type: none"> ▪ Low visual quality output (256×256)
FaR-GAN [92]	GAN	Facial landmark and Boundary features	<ul style="list-style-type: none"> ▪ VGG Face dataset ▪ VoxCeleb1 dataset 	256×256	<ul style="list-style-type: none"> ▪ Sensitive to source identity leakage ▪ Lack of gaze adaption
MarioNETte [43]	GAN (2Encoder-1Decoder)	Facial landmark features	VoxCeleb1	256×256	<ul style="list-style-type: none"> ▪ Fails to preserve source facial characteristics completely
FSGAN[67]	GAN+RNN	<ul style="list-style-type: none"> ▪ Facial landmarks ▪ LFW parts label set 	<ul style="list-style-type: none"> ▪ IJB-C dataset (5500 face videos) ▪ VGGFace2 ▪ CelebA ▪ Figaro dataset 	256×256	<ul style="list-style-type: none"> ▪ 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
Face Synthesis					
Karras et al. [100]	StyleGAN	<ul style="list-style-type: none"> ▪ Deep Features 	<ul style="list-style-type: none"> ▪ ImageNet 	1024×1024	<ul style="list-style-type: none"> ▪ Blob-like artifacts
Huang et al. [101]	TP-GAN	<ul style="list-style-type: none"> ▪ Deep Features 	<ul style="list-style-type: none"> ▪ LFW 	256x256	<ul style="list-style-type: none"> ▪ Lack fine details ▪ Lack semantic consistency
Zhang et al. [102]	SAGAN	<ul style="list-style-type: none"> ▪ Deep Features 	<ul style="list-style-type: none"> ▪ ImageNet2012 	128×128	<ul style="list-style-type: none"> ▪ Unwanted visible artifacts
Brock et al. [103]	BigGAN	<ul style="list-style-type: none"> ▪ Deep Features 	<ul style="list-style-type: none"> ▪ ImageNet 	512×512	<ul style="list-style-type: none"> ▪ Class-conditional image synthesis ▪ Class leakage
Zhang et al. [104]	StackGAN	<ul style="list-style-type: none"> ▪ Deep Features 	<ul style="list-style-type: none"> ▪ CUB ▪ Oxford ▪ MS-COCO 	256×256	<ul style="list-style-type: none"> ▪ Lack semantic consistency
Face attribute Editing					
Perarnau et al. [105]	IcGAN	<ul style="list-style-type: none"> ▪ Deep Features 	<ul style="list-style-type: none"> ▪ CelebA ▪ MNIST 	64×64	<ul style="list-style-type: none"> ▪ Fails to preserve original face identity
Fader Network [106]	Encoder-decoder	<ul style="list-style-type: none"> ▪ Deep Features 	<ul style="list-style-type: none"> ▪ CelebA 	256×256	<ul style="list-style-type: none"> ▪ Unwanted distortion and blurriness ▪ Fails to preserve fine details
Choi et al. [107]	StarGAN	<ul style="list-style-type: none"> ▪ Deep Features 	<ul style="list-style-type: none"> ▪ CelebA ▪ RaFD 	512×512	<ul style="list-style-type: none"> ▪ Undesired visible artifacts in the facial skin e.g., uneven color tone
He et al. [108]	AttGAN	<ul style="list-style-type: none"> ▪ Deep Features 	<ul style="list-style-type: none"> ▪ CelebA ▪ LFW 	384 × 384	<ul style="list-style-type: none"> ▪ Generates low quality results and adds unwanted changes, blurriness
Liu et al. [109]	STGAN	<ul style="list-style-type: none"> ▪ Deep Features 	<ul style="list-style-type: none"> ▪ CelebA 	384×384	<ul style="list-style-type: none"> ▪ Poor performance for multiple attribute manipulation
Zhang et al. [112]	SAGAN	<ul style="list-style-type: none"> ▪ Deep Features 	<ul style="list-style-type: none"> ▪ CelebA 	256×256	<ul style="list-style-type: none"> ▪ Lack of details in attribute-irrelevant region
He et al. [113]	PA-GAN	<ul style="list-style-type: none"> ▪ Deep Features 	<ul style="list-style-type: none"> ▪ CelebA 	256×256	<ul style="list-style-type: none"> ▪ undesired artifacts in case of baldness and open mouth etc.

SAGAN [112] introduced a GAN based framework comprising an 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 [113] employed a progressive attention mechanism in 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 the multiple attributes manipulation and well preserves irrelevance within a single model. However, some undesired artifacts appear in cases where significant modifications are required such as the baldness and open mouth.

4.5 Audio Deepfakes

AI-synthesized audio manipulation is a type of deepfake that can clone a person's voice and depict that voice saying something outrageous, that the person never said. Recent advancements in AI-synthesized algorithms for speech synthesis and voice cloning have shown a 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 utterances of the target speaker, with highly convincing results [50, 114]. The 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 and are being used for malicious purposes, such as political gains, fake news, and fraudulent scams, etc. More complex audio synthesis could be combining power of AI and manual editing. For example, neural network-powered voice synthesis models, such as Google's Tacotron [48], Wavenet [47] or AdobeVoco [115], can generate realistic and convincing fake voices that resemble the victim's voice, as the first step. Later on, audio editing software, e.g. Audacity [4], can be used to combine the different pieces of original and synthesized audios to make more powerful audios.

AI-based impersonation is not limited to visual content; recent advancements in AI-synthesized fake voices are assisting the creation of highly realistic deepfakes videos. Recent developments in speech synthesis have shown their potential to produce realistic and natural audio deepfakes, exhibiting real threats to society [32]. Combining synthetic audio content with visual manipulation can significantly make deepfake videos more convincing and increase their harmful impact [32]. Until now, however, these synthesized speeches lack some aspects of voice quality, like expressiveness, roughness, breathiness, stress, and emotion, etc., specific to a target identity. The AI research community has made some efforts to produce human-like voice quality with high speaker similarity. This section lists the latest progress in speech synthesis and describes the alarming outcomes in speech synthesis and the potential threat to steal a voice identity.

Speech synthesis refers to a technology that can generate speech from a given input i.e., text-to-speech (TTS) [116] or voice conversion (VC) [117]. TTS is a decades-old technology that can synthesize the natural-sounding voice of a speaker from a given input text, and thus enables a voice to be used for better human-computer interaction. VC is another technique that modifies the audio waveform of a source speaker to make it sound like the target speaker's voice while keeping the linguistic content unchanged [118]. The latest speech synthesis initiatives raise more concerns about the reliability of the speech/audio [119].

Overall, important developments in speech synthesis have been done using the methods of speech concatenation or parameterization. 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 and consistent. In contrast, parametric models emphasize extracting acoustic features from the given text inputs and converting them into an audio signal using the vocoders. Interesting outcomes of parametric TTS due to improved speech parameterization performance, vocal tract modeling, and the implementation of deep neural networks evidently show the future of artificial speech production [119]. Fig. 7 shows the principle design of modern TTS methods.

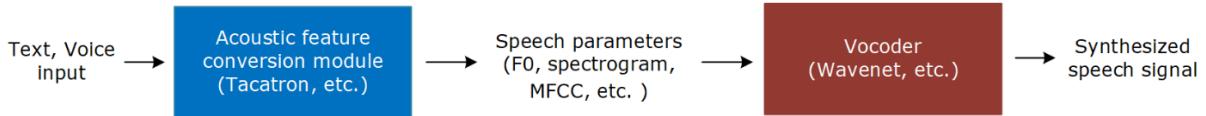


Figure 7: Workflow diagram of the latest parametric TTS systems

Over the last few years, naturalness and quality of TTS systems have improved significantly due to recent advancements in deep learning techniques. The significant developments in voice/speech synthesis are WaveNet [47], Tacotron [48], and DeepVoice3 [120], which can generate realistic sounding synthetic speech from text input to provide an enhanced interaction experience between humans and machines. Table 3 presents an overview of state-of-the-art speech synthesis methods. WaveNet [47] was developed by DeepMind in 2016 and evolved from pixelCNN [117]. WaveNet models utilize raw audio waveforms by using acoustic features, i.e. spectrograms, through a generative framework that is trained on actual recorded speech. WaveNet is a probabilistic autoregressive model that

works by determining the probability distribution of the current acoustic signal by using probabilities of previously generated samples. Dilated causal convolutions are the main modules that are utilized to guarantee that WaveNet can only employ the sampling points from 0 to $t-1$ for predicting a new sampling point. Although the WaveNet framework is capable of generating audio of fine quality it has the following limitations: i) it is a time-consuming process, as the generation of the new signal is dependent on all previously generated samples, and ii) the dependency of WaveNet on linguistic features has a negative impact on the synthesis process. So, to deal with the aforementioned challenges, parallel WaveNet has been introduced to enhance the sampling efficacy. The new model is proficient in producing high-fidelity audio signals [121]. Another DL based using a variant of WaveNet, namely Deep Voice 1 [49], is presented by replacing 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 [48] an end-to-end speech synthesis model. Tacotron can synthesize speech from given <text, audio> pairs and thus generalizes well to other datasets. Similar to WaveNet, the Tacotron framework is a generative framework comprised of a seq2seq model that contains an encoder, an attention-based decoder, and a post-processing network. The framework accepts characters as input and generates a raw spectrogram, which is later transformed into waveforms. The model employs the Griffin–Lim technique [122] to rebuild the acoustic signal by computing phase data through the spectrogram iteratively. Even though the Tacotron model has attained better performance it has one potential limitation i.e. it must employ multiple recurrent components. The inclusion of these units makes it economically inefficient, so that it requires high-performance systems for model training. Deep Voice 2 [123] combines the capabilities of both the Tacotron and WaveNet models for voice synthesis. Initially, Tacotron is employed for converting the input text to a linear scale spectrogram, then it is later converted to voice through the WaveNet model.

In [124], Tacotron2 was introduced for vocal synthesis and it exhibits an impressive 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. The rest of the framework follows a modified WaveNet model which works as a vocoder and generates time-domain signals through spectrograms. To deal with the time complexities of recurrent unit based speech synthesis models, a new, fully-convolutional character-to-spectrogram model named DeepVoice3 [120] was presented. 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 [125], 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 which are later decoded in short vectors. The new context vector is produced by assessing the encoding of the resulting phonemes and summing them together. A few distinguishing properties of VoiceLoop from its peers are an inclusion of a memory buffer as a replacement for the conventional RNNs, shared memory among all procedures, and the employment of shallow, fully connected networks for all processing. These properties make VoiceLoop adaptable for the scenario where the speaker's voice is recorded in a noisy environment. The above mentioned powerful end-to-end speech synthesizer models [120, 124] 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. These classical content-oriented TTS systems are now evolving into a system with personalized voices e.g., a specific subject's speech identity and even a real person's voice cloning. Eventually, the goal will be to improve the unnatural machine expression in human-machine interaction by replacing it with ultra-natural speech. At the same time, voice cloning presents a security risk and may result in identity theft.

The latest TTS systems can convert given text to a human speech with a particular voice identity. This synthesized speech may belong to a particular individual seen during the training process or even a non-existing individual by mixing voices of other identities. To generate a target-specific voice, a speech synthesis system requires retraining of the model on several hours of the recorded target's speech. This limits the ability of technology to scale for many different voices and languages. A voice cloning system needs to adapt an unseen speaker's voice to a generative model learned from scratch on a large dataset. Using generative models, researchers have built voice imitating TTS models that can clone the voice of a particular speaker in real-time using only a few seconds of an available reference speech sample [126]. 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 [127]. Various AI-enabled voice cloning online platforms are available such as

Overdub¹, VoiceApp², and iSpeech³ which can produce synthesized fake voices that closely resemble target speech and gives the public access to this technology.

Jia et al. [126] proposed 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: (i) a recurrent speaker encoder that computes a fixed dimensional feature vector from the input signal, (ii) a Tacotron 2 based sequence-to-sequence synthesizer that predicts a mel-spectrogram from text depending on the embedding vector of the speaker, and (iii) a WaveNet [47] based neural vocoder that translates the spectrogram into time-domain waveforms. 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. [50] proposed a Deep Voice 3 based voice cloning system that can generate the cloned voice of any target by using a small number of recorded audio samples. This technique [50] is 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. To clone the speaker's voice the model computes the characteristics of the speaker to produce the cloned audio signal provided through the given text with an average cloning time of only 3.7 seconds.

Table 3: An overview of the state-of-the-art speech synthesis techniques

Methods	Technique	Features	Dataset	Limitations
WaveNet [47]	Deep neural network	▪ linguistic features ▪ fundamental frequency (log F0)	▪ VCTK (44 hrs.)	▪ Computationally expensive
Tacotron [48]	Encoder-Decoder with RNN	▪ Deep features	Private (24.6 hrs.)	▪ Costly to train the model
Deep Voice 1[49]	Deep neural networks	▪ linguistic features	Private (20 hrs.)	▪ Independent training of each module leads to a cumulative error in synthesized speech
Deep Voice 2 [123]	RNN	▪ Deep features	VCTK (44 hrs.)	▪ Costly to train the model
DeepVoice3 [120]	Encoder-decoder	▪ Deep features	▪ Private (20 hrs.) ▪ VCTK (44 hrs.) ▪ LibriSpeech ASR (820 hrs.)	▪ Does not generalized well for unseen samples.
Parallel WaveNet [121]	Feed-forward neural network with dilated causal convolutions	▪ linguistic features	Private	▪ Requires a large amount of target's speech training data.
VoiceLoop [125]	Fully-connected neural network	▪ 63-dimensional audio features	▪ VCTK (44 hrs.) ▪ Private	▪ Low ecological validity
Tacotron2[124]	▪ Encoder-decoder	▪ linguistic features	Japanese speech corpus from the ATR Ximera dataset (46.9 hrs.)	▪ Lack of real time speech synthesis
Arik et al. [50]	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. [126]	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. [127]	Encoder-decoder	▪ Mel spectrograms	▪ LibriSpeech (245 hrs.) ▪ VCTK (44 hrs.)	▪ Low performance in the case of noisy audio samples
Chen et al. [128]	Encoder + deep neural network	▪ Mel spectrograms	▪ LibriSpeech (820 hrs.) ▪ private	▪ Low performance in the case of a low-quality audio sample
Cong et al. [129]	Encoder-decoder	▪ Mel spectrograms	▪ MULTI-SPK ▪ CHiME-4	▪ Lacks in synthesizing utterances of a target speaker

Luong et al. [127] proposed a voice cloning framework that can synthesize 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

¹ <https://www.descript.com/overdub>

² <https://apps.apple.com/us/app/voiceapp/id1122985291>

³ <https://www.ispeech.org/apps>

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 generate the cloned speech. Chen et al. [128] proposed a meta-learning approach using 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 speakers. This approach performs well on clean and high-quality training data.

Currently voice cloning systems produce quality speech through speaker adaptation or speaker encoding from a clean speech sample. The presence of noise deviates the speaker encoding and directly affects the performance of synthesized speech. In [129], the authors proposed a seq2seq multi-speaker framework with domain adversarial training to produce a target speaker voice from only a few available noisy samples. The results showed improved naturalness of synthetic speech. However, similarity still remains challenging due to lack of transferring target accents, and prosody to synthesized speech with a limited amount of low-quality speech data.

4.6 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 driving content 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 driving identity, specifically in face reenactment tasks where target expressions are driven by some source identity. The facial data of the driving 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 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 items. 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 jitter 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.

Lack of realism in synthetic audio: Though the quality is certainly getting much better, there is still a need for improvement. The main challenges of audio-based deepfakes are the lack of natural emotions, pauses, breathiness, and the pace at which the target speaks.

Based on the above-mentioned limitations we can argue that there exists a need to develop effective deepfake generation methods that are robust to variations in illumination conditions, temporal coherence, occlusions, pose variations, camera distance, identity leakage, and paired training.

5 Deepfakes detection techniques

The evolution of ML and the emergence of advanced artificial intelligence algorithms have increased the ease with which fake multimedia content is producing images, audio, or videos, and has improved the realism of manipulated information dramatically [8, 86, 130-132]. It is very difficult now for people to differentiate between actual and synthesized multimedia (image, audio, video, etc.). Deepfakes have the potential to initiate political tension, conflicts, violence, and war worldwide. This results in a violation of privacy and poses a serious threat to societal security and democracy. Therefore, to overcome the devastating effects of deepfakes, multimedia forensic techniques for deepfake detection has grasped the attention of researchers. Existing approaches have either targeted spatial and temporal artifacts left during generation, or data-driven classification (Fig. 8). The spatial artifacts include inconsistencies [133-139], abnormalities in background [140-142], and GAN fingerprints [143-145]. The temporal artifacts involve detecting variation in a person's behavior [146, 147], physiological signals [135, 148-150], coherence [151, 152], and video frame synchronization [153-156]. Instead of focusing on a specific artifact, some approaches are data-driven, which detect manipulations by classification [157-177] or anomaly identification [178-182]. Fig. 9 shows a general deepfakes detection process pipeline. For feature extraction, all deepfake detection approaches have employed either handcrafted features-based or deep learning-based methods. We have discussed both types of methods in the subsequent sections.

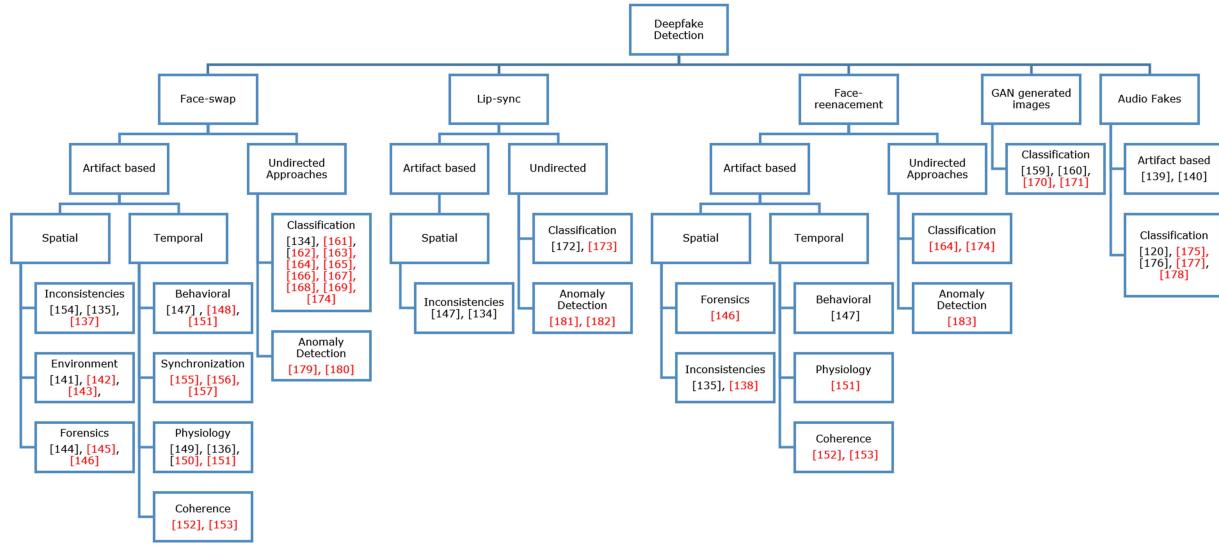


Figure 8: Categorization of deepfake detection techniques (The red color shows deep learning-based technique and black is for handcrafted techniques)

5.1 Handcrafted feature-based techniques

A lot of literature is available on image and video forgery detection [183-186]. As AI-manipulated data is a new phenomenon, there are a small number of forensic techniques that work well for deepfake detection. An overview of deepfake detection techniques based on handcrafted features is presented in Table 4. Recently, some researchers [157, 187] have adopted the idea of employing the traditional methods of image forgery identification to deepfake detection. Zhang et al. [157] proposed a technique to detect swapped faces. SURF descriptor was employed on the images for feature extraction that were then used to train the SVM for classification. This technique was then tested on the set of Gaussian blurred images. This approach has improved the deepfakes detection performance but has two potential limitations. Firstly, this approach is unable to preserve the facial expression of the given image. Secondly, this technique only works on still images and is unable to detect manipulated videos. Yang et al. [143] introduced an approach to detect the deepfakes by estimating the 3D head position from 2D facial landmarks. The computed difference among the head poses was used as a feature vector to train the SVM classifier and was 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 the blurred images, which degrades the performance of this method under such scenarios. Korshunov et al. [158] employed the Image Quality Metric features along with principal component analysis and linear discriminant analysis (LDA) for feature extraction and then trained the SVM to classify the video content as bonafide or fake. It is concluded from [158] that existing face recognition techniques like Facenet [188] and Visual Geometry Group (VGG) [189] are unable to detect deepfakes. Moreover, pure lip-syncing based

approaches are unable to detect the GAN generated videos. The SVM classifier exhibits better deepfake detection performance; however, they do not perform well for high-quality visual contents.

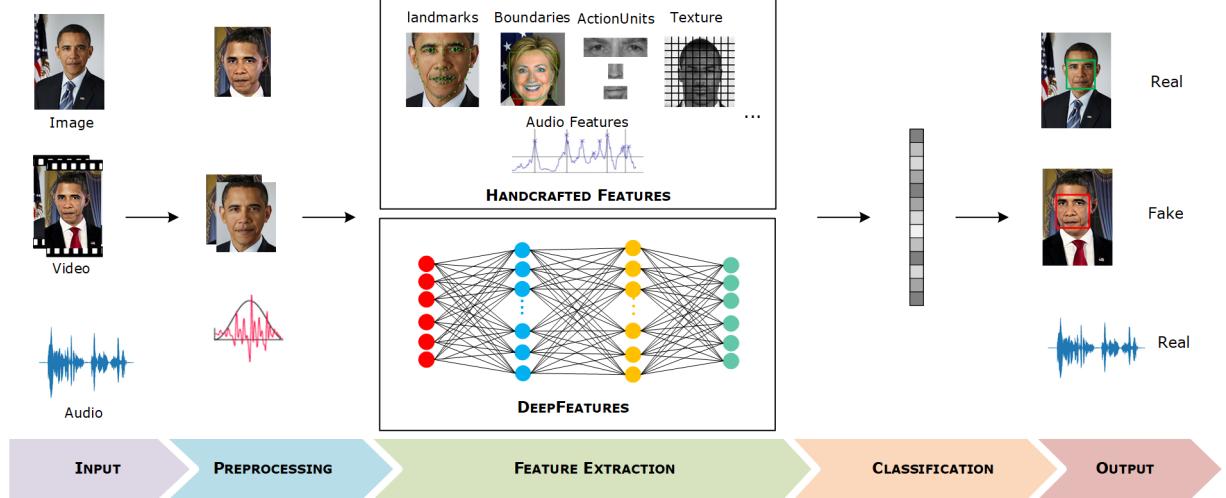


Figure 9: The general processing pipeline of deepfake detection

Agarwal et al. [146] presented a user-specific technique for deepfakes detection. First, GAN was used to generate all three types of deepfakes for US ex-president Barack Obama. Then the OpenFace2 [190] 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 the binary SVM to classify between the original face and synthesized face of Barack Obama. This technique provides good detection accuracy for face swap and puppet master over lip-syncing, however, it is vulnerable in those scenarios where a person is looking off-camera. Guera et al. [153] presented a method for detecting synthesized faces from videos. Multimedia stream descriptors [191] were used to extract the features that were then used to train the SVM, and random forest classifiers to differentiate between the real and manipulated faces from the videos. This technique gives an effective solution to deepfakes detection but is not applicable to video re-encoding attacks.

Korshunov et al. [133] proposed a technique to detect lip-sync-based deepfakes. The 40-D MFCC features containing the 13-D MFCC, 13-D delta, and 13-D double-delta, along with the energy, were used in combination with mouth landmarks to train the four classifiers, i.e. SVM, LSTM, multilayer perceptron (MLP), and Gaussian mixture model (GMM). Three publicly available datasets, named VidTIMIT [192], AMI corpus [193], and GRID corpus [194] were used to evaluate the performance of this technique. From the results, it was concluded in [133] that LSTM achieves better performance over other techniques. However, lip-syncing deepfake detection performance of the LSTM method drops for the VidTIMIT [192] and AMI [193] datasets due to fewer training samples for each person in both of these datasets over the GRID dataset. Ciftci et al. [148] introduced an approach to detect forensic changes within videos by computing the biological signals (e.g. heart rate) from the face portion of the videos. Temporal and spatial characteristics of facial features were computed to train the SVM and CNN model to differentiate between bona fide 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. Matern et al. [134] 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 the manipulated content from the original data. This technique has a low computational cost; however, it is applicable only to the visual content with open eyes or visible teeth. Jung et al. [135] proposed 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 [195] and EAR technique (eye detect) [196] to detect eye blinking. An integrity authentication method was employed by tracking the fluctuation of eye blinks based on gender, age, behavior, and time factor to spot the real and fake videos. The approach in [135] exhibits better deepfake detection performance, however it is not appropriate if subject in the video is suffering from mental illness as we often experience abnormal eye blinking pattern for such people. Guarnera et al. [159] presented an approach to detect the image manipulation. Initially, Expectation-Maximization (EM) technique was applied to obtain the image features based on which naive classifier was trained to discriminate against

original and fake images. This approach shows better deepfake identification accuracy, however, it is only applicable to static images.

Table 4: An overview of Deepfake detection techniques based on handcrafted features and their limitations

Author	Technique	Features	Best Evaluation performance	Dataset	Limitations
Face-swap					
Zhang et al. [157]	SURF + SVM	64-D features using SURF	<ul style="list-style-type: none"> ▪ Precision= 97% ▪ Recall= 88% ▪ Accuracy= 92% ▪ 	Generate deepfake dataset using LFW face database.	<ul style="list-style-type: none"> ▪ Unable to preserve facial expressions ▪ Works with static images only.
Yang et al. [143]	SVM Classifier	68-D facial landmarks using DLib	<ul style="list-style-type: none"> ▪ ROC=89% ▪ ROC=84% 	<ul style="list-style-type: none"> ▪ UADFV ▪ DARPA MediFor GAN Image/ Video Challenge. 	Degraded performance for blurry images.
Guera et al. [153]	SVM, RF Classifier	Multimedia stream descriptor [29]	<ul style="list-style-type: none"> ▪ AUC= 93% (SVM) ▪ AUC= 96% (RF) 	Custom dataset.	Fails on video re-encoding attacks
Ciftci et al. [148]	CNN	medical signals features	▪ Accuracy= 96%	Face Forensics dataset	Large feature vector space.
Jung et al. [135]	Fast-HyperFace[195], EAR[196]	Landmark features	▪ Accuracy= 87.5%	Eye Blinking Prediction dataset	Inappropriate for people with mental illness
Lip-sync					
Korshunov et al. [133]	SVM, LSTM, MLP, GMM	MFFC + mouth landmark features	<ul style="list-style-type: none"> ▪ 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) 	<ul style="list-style-type: none"> ▪ VidTIMIT ▪ AMI ▪ GRID 	LSTM performs better than others but its performance degrades as the training samples decrease.
Face-swap (FS) & Face Reenactment (FR)					
Matern et al. [134]	MLP, Logreg	16-D texture energy based features of eyes and teeth [197]	<ul style="list-style-type: none"> ▪ AUC=.0.851(MLP-AUC for FS) ▪ AUC=0.784 (LogReg-AUC for FS) ▪ AUC=.823 (MLP-AUC for FR) ▪ AUC=.866 (LogReg-AUC for FR) 	FaceForensics++	Only applicable to face images with open eyes and clear teeth.
All types					
Agarwal et al. [146]	SVM Classifier	16 AU's using OpenFace2 toolkit	<ul style="list-style-type: none"> ▪ AUC= 93% (FS) ▪ AUC= 95% (lip-sync) ▪ AUC= 98% (FR) 	Own dataset.	Degraded performance in cases where a person is looking off-camera.
GAN generated Fake images					
Guarnera et al. [159]	EM + (KNN, SVM, LDA)	Deep features	<ul style="list-style-type: none"> ▪ Accuracy=99.22 (KNN) ▪ Accuracy= 99.81(SVM) ▪ Accuracy= 99.61 (LDA) 	CelebA	Not robust to compressed images.
McCloskey et al. [140]	SVM	Color channels	▪ AUC=70%	MFC2018	Performance degrades over blurry samples.
Audio Manipulation					
Nagarshetha et al. [198]	SVM	HFCC, CQCC	EER of 11.5%	ASVspoof 2017	<ul style="list-style-type: none"> ▪ Does not generalize well to different classes of spoofing attacks
Gunendradasan et al. [199]	GMM	TLC-AM TLC-FM	EER=8.68(TLC-AM), 11.30 (TLC-FM)	ASV spoof-2017	Computationally complex
Witkowski et al. [200]	GMM	CQCC, Cepstrum, IMFCC,MFCC, LPCCres	EER=5.13 (CQCC), 3.38(Cepstrum), 4.16 (IMFCC), 16.76(MFCC), 6.37(LPCCres)	ASVspoof 2017	Computationally complex
Saranya et al. [201]	GMM	MFCC, CQCC, and MFS	EER=19.36	ASVspoof 2017	Requires performance improvement

McCloskey et al. [140] presented an approach to identify fake images by employing the fact that the color information is evidently dissimilar between real camera and fake synthesis samples. The color key-points from input samples were used to train the SVM for classification. This approach [140] exhibits better fake sample detection accuracy, however, it may not perform well for blurred images. Guarnera et al. [159] proposed a method to identify fake images. Initially, the EM algorithm was used to calculate the image features. The computed key-points were used to train three types of classifiers, KNN, SVM, and LDA. The approach in [159] performs well for synthesized image identification, but may not perform well for compressed images.

Nagarsheth et al. [198] presented an approach to identify audio replay spoofing. Initially, high-frequency cepstral coefficients (HFCC) and CQCC features were used to create embeddings through a DNN. Next, an SVM was trained over computed embeddings for replay attack identification. This approach [198] exhibits better audio manipulation detection accuracy, however may not generalize well to different classes of spoofing attacks. A few works [199, 200] have stated the significance of high-frequency band analysis to better detect features present in replay audio. In [199], transmission line cochlea-amplitude modulation (TLC-AM) and TLC-frequency modulation were employed to train a GMM for replay spoofing identification. Similarly in [200], inverted-MFCC (IMFCC), linear predictive cepstral coefficients (LPCC), and LPCCres high-frequency band attributes were utilized along with spectral features, i.e. CQCC, MFCC, and Cepstrum to train the GMM for replay detection. Though the approaches in [199, 200] exhibit better performance over the ASVspoof baseline model it is with the overhead of increased feature computation cost. In [201], MFCC, CQCC, and Mel-Filterbank-Slope (MFS) features were utilized with the GMM to detect replay attacks.

5.2 Deep learning-based techniques

Handcrafted feature-based techniques have been frequently used for identifying manipulated content. These approaches often work well for detecting the forensic changes made within static digital images but may not perform well for deepfakes due to the following reasons: i) temporal characteristics of videos vary from frame to frame, and ii) compression techniques are usually applied after altering the information in videos so very important visual information is dropped, which degrades the performance of these techniques [163]. Therefore, to overcome the problems of the handcrafted feature-based techniques, Deep Learning (DL) approaches are being heavily explored these days. A summary of deepfake detection techniques based on deep learning approaches can be found in Table 5. Li et al. [160] proposed a method of detecting the forensic modifications made within the videos. First, the facial landmarks were extracted using the dlib software package [202]. Next, CNN based models named ResNet152, ResNet101, ResNet50, and VGG16 were trained to detect forged content from videos. This approach is more robust in detecting the forensic changes; however, it exhibits low performance on multi-time compressed videos. Guera et al. [154] proposed a technique for deepfakes detection. CNN was used to extract the features at the frame level. Then the RNN was trained on the set of extracted features to detect deepfakes from the input videos. This work achieves good detection performance but only on videos of short duration i.e. videos of 2 seconds or less.

Li et al. [149] proposed a technique to detect deepfakes by using the fact that the manipulated videos lack accurate eye blinking in synthesized faces. CNN/RNN approach was used to detect the lack of eye blinking in the videos 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 the 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. Nataraj et al. [161] 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. Sabir et al. [155] observed that while generating the manipulated content, forgers often do not impose temporal coherence in the synthesis process. So, in [155], a recurrent convolutional model was used to investigate the temporal artifacts to identify synthesized faces in the images. These techniques [155, 161] achieve better detection performance, however they can only process static images.

Rossler et al. [162] employed both the handcrafted (co-occurrence matrix) and learned features for detecting manipulated content. It was concluded in [162] that the detection performance of both networks, either employing hand-crafted or deep features, degrade when evaluating them on compressed videos. To analyze the mesoscopic properties of manipulated content, Afchar et al. [163] 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 has managed to reduce the computational cost by down sampling the frames, but at the expense of a decrease in accuracy in deepfake detection. Nguyen et al. [164] proposed a multi-task, learning-based CNN network to simultaneously detect and localize manipulated content from the 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 is robust to deepfakes detection; however, the evaluation accuracy degrades over unseen scenarios. To overcome the issue of performance degradation as in [164], Stehouwer et al. [165] proposed a Forensic transfer (FT) based CNN approach for deepfake detection. This work [165], however, suffers from high computational cost due to a large feature space. Amerini et al. [156] proposed an approach based on optical flow fields to detect synthesized faces in digital videos. The optical flow fields [203] of each video frame were computed using PWC-Net [204]. The estimated optical flow fields of frames were used to train the VGG16 and ResNet50 to classify bonafide and fake content. This method [156] exhibits better deepfake detection performance, however, only initial results have been reported.

Montserrat et al. [166] introduced a method of locating forensic manipulations made within videos. Initially, MTCNN [205] was employed to detect the faces from all video frames on which CNN was applied, to compute the features. In the next step, the Automatic Face Weighting (AFW) mechanism, along with a Gated Recurrent Unit, was used to discard the false-detected faces. Finally, an RNN was employed to combine the features from all steps and locate the manipulated content in the videos. The approach in [166] works well for deepfake detection, however, it is unable to obtain the prediction from the features in multiple frames. Lima et al. [167] introduced a technique to detect video manipulation by learning the temporal information of frames. Initially, VGG-11 was employed to compute the features from video frames, on which LSTM was applied for temporal sequence analysis. Several CNN frameworks, named R3D, ResNet, I3D, were trained on the temporal sequence descriptors outputted by the LSTM, to identify original and manipulated videos. This approach [167] improves deepfake detection accuracy but at the expense of high computational cost. Agarwal et al. [147] presented an approach to locate face-swap-based manipulations by combining both facial and behavioral biometrics. The behavioral biometric was recognized with the encoder-decoder network (Facial Attributes-Net, FAB-Net) [206]. Whereas VGG-16 was employed for facial features computation. Finally, by merging both metrics the inconsistencies in the matching identities were revealed to locate face-swap deepfakes. This approach [147] works well for unseen cases, however, it may not generalize well to lip-synch-based deepfakes. Fernandes et al. [150] introduced a technique to locate visual manipulation by measuring the heart-rate of the subjects. Initially, three techniques: skin color variation [207], average optical intensity [208], and Eulerian video magnification [209], were used to measure heart rate. The computed heart-rate was used to train a Neural Ordinary Differential Equations (Neural-ODE) model [210] to differentiate the original and altered content. This technique [150] works well for deepfakes detection but has increased computational complexity.

Wang et al. [178] introduced a technique to locate synthesized faces. Initially, deep-features were computed by employing VGG-Face [211] and used to train an SVM for fake-faces classification. The approach in [178] works well under the presence of compression operations but its performance degrades significantly for additive noise attacks. Yu et al. [144] 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 [144] exhibits better detection accuracy, however, it may not perform well with post-processing operations i.e. noise, compression, and blurring, etc. Marra et al. [168] proposed a study to identify the GAN-generated fake images. Particularly, [168] 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 [212], named as Multi-Task MultiClassifier, and Multi-Task Single Classifier. This approach [168] is robust to unseen GAN generated samples but unable to perform well if information on the fake content generation method is not available.

Chen et al. [213] introduced a technique to identify audio manipulation. The approach [213] works by employing a large margin cosine loss function (LMCL) along with online frequency masking augmentation to train the NN to learn more discriminative key-point embedding. This technique [213] shows better audio manipulation detection accuracy but may not perform well in the presence of noisy conditions. Huang et al. [214] presented an approach for audio spoofing detection. Initially, short-term zero-crossing rate and energy were utilized to identify the silent segments from each speech signal. In the next step, the linear filter bank (LFBank) key-points were computed from the nominated segments in the relatively high-frequency domain. Lastly, an attention-enhanced DenseNet-BiLSTM framework was built to locate audio manipulations. This method [214] can avoid the over-fitting, however, it is at the expense of high computational cost. Wu et al. [177] introduced a novel key-points genuinization based light convolutional neural networks (LCNN) framework for the identification of synthetic speech manipulation. The attributes of the original speech were utilized to train a model using a CNN. It was then converted to an original key-point distribution closer to that of genuine speech. The transformed key-points were used with an LCNN to identify genuine and altered speech. This approach [177] is robust to synthetic speech manipulation detection. It is, however, unable to deal with replay attack detection. Lai et al. [215] presented an approach for locating voice manipulation. Initially, the Attentive Filtering Network was employed for key-point engineering, based on which ResNet-based

Classifier was trained to detect the replay attacks. The approach in [215] is robust to speech alteration detection, however, performance can be further improved. Yang et al. [216] employed inverted constant-Q coefficients, inverted constant-Q cepstral coefficients, constant-Q block coefficients, and inverted constant-Q linear block coefficients to train a DNN to differentiate between actual and spoofed speech. This approach [216] is robust to noisy conditions, however, it is unable to perform well for real-world scenarios.

Table 5: An overview of Deepfakes detection techniques based on deep learning and their limitations

Author	Technique	Features	Best Evaluation performance	Dataset	Limitations
Face-swap					
Li et al. [160]	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. [154]	CNN/ RNN	Deep features	Accuracy=97.1%	Customized dataset	▪ Applicable to short videos only (2 sec).
Li et al. [149]	CNN/RNN	DLib facial landmarks	TPR= 99%	Customized dataset	▪ Fails over frequent and closed eyes blinking.
Montserrat et al. [166]	CNN + RNN	Deep features	Accuracy=92.61%	DFDC	▪ Performance needs improvement.
Lima et al. [167]	VGG11 + LSTM	Deep features	Accuracy= 98.26%, AUC= 99.73%	Celeb-DF	▪ Computationally complex.
Agarwal et al. [147]	VGG6 + encoder-decoder network	Deep features + behavioral biometrics	AUC= 99%	WLDR	▪ May not generalize well to lip-synch based deepfakes.
			AUC= 99%	FF	
			AUC= 93%	DFD	
			AUC= 99%	Celeb-DF	
Fernandes et al. [150]	Neural-ODE model	Heart-rate	Loss=0.0215	Custom	▪ Computationally expensive
			Loss=0.0327	DeepfakeTIMIT	
Face Reenactment					
Amerini et al. [156]	VGG16, ResNet	Optical flow fields	Accuracy= 81.61% (VGG16), 75.46% (ResNet)	FaceForensic++	▪ Very few results are reported
Face-swap (FS) & Face Reenactment (FR)					
Sabir et al. [155]	CNN/RNN	CNN features	Accuracy= 96.3% (FS), 94.35 % (FR)	FaceForenciss++	▪ Results are reported for static images only.
Afchar et al. [163]	MesoInception-4	Deep features (DF)	TPR= 81.3 % (FS) TPR= 81.3% (FR)	FaceForenciss++	▪ Performance degrades on low quality videos.
Nguyen et al. [164]	CNN	Deep features	Accuracy=83.71% (FS), 92.50% (FR)	FaceForensics++	▪ Degraded detection performance for unseen cases.
Stehouwer et al. [165]	CNN	Deep features	Accuracy=99.43% (FS), 99.4% (FR)	Diverse Fake Face Dataset (DFFD)	▪ Computationally expensive due to large feature vector space.
Rossle et al. [162]	SVM + CNN	Co-Occurance matrix + DF	Accuracy= 90.29% (FS), 86.86% (FR)	FaceForensics++	▪ Low performance on compressed videos.
GAN-generated fake images					
Nataraj et al [161]	CNN	Deep features + co-occurrence matrices	Accuracy = 99.49%	▪ cycleGAN	▪ Works with static images only. ▪ Low performance for jpeg compressed images.
			Accuracy = 93.42%	▪ StarGAN	
Yu et al. [144]	CNN	Deep features	Accuracy = 99.43%	CelebA	▪ Poor performance on post-processing operations.
Marra et al. [168]	CNN + Incremental Learning	Deep features	Accuracy = 99.3%	Customized	▪ Needs source manipulation technique information
Audio manipulation					
Chen et al. [213]	LMCL	60-D LFBank	EER= 1.26%	ASVspoof 2019 challenge [217]	▪ Not robust to noisy conditions.
Huang et al. [214]	DenseNet-BiLSTM	LFBank	EER= 6.43%	▪ BTAS2016 [218]	▪ Computationally complex approach.
			EER=0.53%	▪ ASVspoof 2019 challenge [217]	
Wu et al. [177]	LCNN	genuine speech features	EER= 4.07%	ASVspoof 2019 challenge	▪ Can't deal with replay attack detection.
Lai et al. [215]	ResNet	time-frequency maps	EER= 8.99%	ASVspoof 2017 challenge [219]	▪ Performance needs improvement.

5.3 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.

5.3.1 Quality of Deepfake Datasets

The accessibility of large databases of deepfakes is an important factor in the generation of deepfake detection techniques. However, analyzing the quality of videos from these datasets reveals several ambiguities in comparison 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, lightning effect, etc., vi) sensitive 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 techniques. Furthermore, manipulated content of low quality can be barely 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 wild.

5.3.2 Performance Evaluation

Presently, deepfake detection methods are formulated as a binary classification problem, where each sample can be either real or fake. Such classification is 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 content 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/visual, and therefore a single label may not be completely accurate. Moreover, in visual content with multiple people's faces, usually one or more of them are manipulated with deepfakes over a segment of frames. Therefore, the binary classification scheme should be enhanced to multiclass/multi-label and local classification/detection at the frame level, to cope with the challenges of real-world scenarios.

5.3.3 Lack of 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 in order for the analysis to be believed before publication or utilization in a court of law. Most deepfakes detection methods lack such an explanation however, particularly those which are based on DL approaches due to their black-box nature.

5.3.4 Temporal Aggregation

Existing deepfake detection methods are based on binary classification at the frame level, i.e. checking the probability of each video frame as real or manipulated. However, these approaches do not consider temporal consistency between frames, 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.

5.3.5 Social Media Laundering

Social platforms like Twitter, Facebook, or Instagram are the main online networks used to spread the audio-visual content among the population. To save the bandwidth of the network or to secure the user's privacy, such content is 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 employing signal level key-points 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.

6 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 over cross datasets to show their generalization power. Therefore, researchers have put in significant effort over the years by preparing the standard

datasets for manipulated visual and audio content. In this section, we have presented a detailed review of the standard datasets that are currently used to evaluate the performance of audio and video deepfake detection techniques. Tables 6 and 7 show a comparison of available video and audio deepfake datasets respectively.

6.1 Video Deepfake datasets

6.1.1 UADFV

The first dataset released for deepfake detection was UADFV [143]. It consists of a total of 98 videos, where 49 are real videos collected from YouTube and manipulated by using the FakeApp application [38] 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 videos is very low, and the resultant alteration is obvious and thus easy to detect.

6.1.2 DeepfakeTIMIT

DeepfakeTIMIT [158] is another standard dataset for deepfake detection which was introduced in 2018. This dataset consists of a total of 620 videos of 32 subjects. For each subject, there are 20 deepfake videos of two quality levels, where 10 videos belong to DeepFake-TIMIT-LQ and the remaining 10 belong to 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 face swap-GAN [65], however, 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.

6.1.3 FaceForensics++

One of the most famous datasets for deepfake detection is FaceForensics++ [162]. This dataset was presented in 2019 as an extended form of the FaceForensics dataset [220], which contains videos with facial expressions manipulation only, and which was released in 2018. The FaceForensics++ dataset has four subsets named FaceSwap [221], DeepFake [39], Face2Face [33], and NeuralTextures [222]. It contains 1000 original videos collected from the YouTube-8M dataset [223] and 3,000 manipulated videos generated using the computer graphics and deepfake approaches specified in [220]. This dataset is also available in two quality levels i.e. uncompressed and H264 compressed format, which can be used to evaluate the performance of deepfake detection approaches on both compressed and uncompressed videos. The FaceForensics++ dataset fails to generalize lip-sync deepfakes however, and some videos exhibit color inconsistencies around the manipulated faces.

6.1.4 Celeb-DF

Another popular dataset used for evaluating deepfake detection techniques is Celeb-DF [141]. 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, which 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 [224, 225].

Table 6: Comparison of Deepfakes detection datasets

	UADFV [143]	DeepFake-TIMIT[158]	FaceForensics++ [162]	Celeb-DF [141]	DFDC [58]
Released	Nov, 2018	Dec, 2018	Jan, 2019	Nov, 2019	Oct, 2019
Total videos	98	620	4000	1203	5250
Real content	48	Nill	1000	408	1131
Fake content	48	620	3000	795	4119
Tool/ technology used for fake content generation	FakeApp application [38]	faceswap- GAN [65]	deepfake, CG-manipulations	deepfake	Unknown
Length	11.4 sec	4 sec	-	13 sec	-
Resolution	294×500	64×64 (LQ) 128×128 (HQ)	480p, 720p, 1080p	various	180p – 2160p
Format	-	JPG	H.264, CRF=0, 23, 40	MPEG4	H.264
Visual quality	low	low	low	high	high
Temporal flickering	yes	yes	yes	improved	improved
modality	visual	Audio/visual	visual	visual	Audio/visual

6.1.5 Deepfake Detection Challenge (DFDC)

Recently, the Facebook community launched a challenge, aptly named the Deepfake Detection Challenge (DFDC) [58], and released a new dataset that contains 1131 original videos and 4119 manipulated videos. The altered content

is generated using two unknown techniques. The DFDC database is publicly available on the Kaggle competition [58]. It contains 100,000 fake videos along with 19,000 original samples.

All of the above-mentioned datasets contain a synthesized face portion only and the datasets lack upper/full body deepfakes. A more robust dataset is needed which should be able to synthesize the entire body of the source person.

6.2 Audio Deepfake Datasets

6.2.1 LJ speech and M-AILabs dataset

LJSpeech [226] and M-AILabs [227] dataset are famous for the real-speech database employed in numerous TTS applications, i.e. DeepVoice 3 [120]. The LJSpeech database is comprised of 13,100 clips totaling 24 hours length. All utterances are recorded by a female speaker. The M-AILABS dataset consists of total 999 hours and 32 minutes of audio. This dataset was created with multiple speakers in 9 different languages.

6.2.2 Mozilla TTS

Mozilla Firefox a well-known publicly available browser, released the biggest open-source database of people speaking [228]. Initially, the database included 1400 hours of recorded voices, in 18 different languages, in 2019. Later it was extended to 7,226 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.

6.2.3 ASV spoof 2019

Another well-known dataset for fake audio detection is ASVsspoof-2019 [229], 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.

6.2.4 Fake-or-Real (FOR) dataset

The FOR database [230] 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[120] and Google-Wavenet [47]) 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-2sec (F2S), and for-rerec (FR). FO contains unbalanced voices without alterations, 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 invader passes a sample via a voice channel (i.e. a cellphone call or a voice message).

6.2.5 Baidu Dataset

The Baidu Silicon Valley AI Lab cloned audio dataset is another database employed for cloned speech detection [50]. This database is comprised of 10 ground truth speech recordings, 120 cloned samples, and 4 morphed samples.

Table 7: Comparison of audio fakes detection datasets

	LJ speech dataset [226]	M-AILabs dataset [227]	Mozilla TTS [228]	FOR dataset [230]	Baidu Dataset [50]	ASV spoof 2019[229]
Released	2017	2019	2019	2019	2018	2019
Total samples	13100	-	5.5 million	195,000	120	122157
Length (hrs)	24	999 hrs 32min	7226	-	0.6	-
Speaker Accent	Native	Native	24% US English, 8% British English	Native	US English, British English	Native
Languages	1	9	54	1	1	1
Speaker gender	100% Female	Male, female	47% Male 15% Female	50% male, 50% female	50% male, 50% female	43% male, 57 female
Format	wav	wav	mp3	mp3	mp3	mp3
Tool/ technology used for generation	recorded	recorded	recorded	Deep Voice 3, TTs, Google-Wavenet etc. [230]	Neural voice cloning [50]	Tacotron2 [9] and WaveNet [10]

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 [231], ii) quick learning, which allows identity stealing using a single image [90, 92], iii) enhancements in visual details [85, 100], iv) improved temporal coherence in generated videos by employing optical flow estimation and GAN based temporal discriminators [74], v) the alleviation of visible artifacts around face boundary by adding secondary networks for seamless blending [232], and vi) improvements in synthesized face quality by adding multiple losses with different responsibilities, such as occlusion, creation, conversion, and blending [79]. 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 manipulations are 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 in 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. The current 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 [75, 233]. 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

Recent 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 [143], lack of eye blinking [196], color variations in facial texture [140] and teeth alignment, ii) detection of unseen GAN generated samples, iii) spatial-temporal features, and iv) psychological signals like heart rate [150], and an individual's behavior patterns [146]. Although extensive work has been presented for automated detection, there is still need for improvement.

- 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.
- Anti-forensic techniques can be employed to mark an original video as a deepfake through the addition of simulated signal level key-points utilized by existing identification methods, a state we named fake deepfake.
- Furthermore, to prevent deepfakes, some authors presented approaches to identify forensic changes made within visual content by employing the concept of blockchain and smart contracts [234, 235]. In [235] 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 the hashes of the interplanetary file system to save videos together with their metadata. This method may perform well for deepfake identification; however, it is applicable only if the metadata of videos do exist. Physics of AI to detect deepfakes using small datasets: The current deepfake detectors face challenges, particularly due to incomplete, sparse, and noisy data in 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 explanatory.

- Existing deepfake detectors have mainly relied on the fixed features of existing cyber-attacks by using ML techniques, including unsupervised clustering and supervised classification methods, and therefore they are less likely to detect unknown deepfakes. Hence, in the future, reinforcement learning (RL) techniques could play a pivotal role in the detection of deepfakes.
- Reinforcement Learning, combined with the Game theory, for detection and to counter antiforensics attacks: 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, thus 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 the deepfake detector 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 [236].
- Since many complex deepfakes comprise temporal sequences of dynamic behaviors, approaches such as [237] could be used to model a detection problem to a state value prediction task of Markov chains. The linear temporal difference (TD) RL algorithm [238] could be used as the state value prediction model, where its outcomes could be compared with a predetermined threshold to distinguish bona fide and deepfake artifacts. Alternatively, the kernel-based RL approach using least-squares TD [239] could also be used. By using kernel methods, the generalization capability of the TD RL is enhanced, especially in high dimensional and nonlinear feature spaces. Therefore, the kernel least squares TD algorithm could be used to predict anomaly probabilities accurately, which would contribute to improving a deepfake detector's performance.
- Hybrid signature, anomaly, and reinforcement learning-based approaches: 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 classify a bona fide multimedia artifact whose patterns are rare in the dataset as an anomaly. On the other hand, signature-based approaches cannot discover unknown attacks [240]. Therefore, the hybrid approach of using both anomaly and signature-based detection needs to be tried out to identify known and unknown attacks. Furthermore, a collaboration with the 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 of deepfakes that contain (or do not contain) anomalies, or any signs of manipulation.
- Feature and classifier fusion: 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 landmark, 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 to improve the accuracy of deepfakes detectors.
- Unified detector to detect multiple forgeries: 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, is 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. replay, cloning, and cloned-replay) and visual forgeries (face-swap, lip-sync, and puppet-master) simultaneously.
- Existing deepfakes datasets lack the potential attributes (i.e. multiple visual and audio forgeries, etc.) required to evaluate the performance of more robust deepfake detection methods. The research community has ignored the fact that deepfake videos contain not only visual forgeries but audio manipulation as well. Existing deepfake datasets do not consider audio forgery and only focus on visual forgeries. In near future, the role of cloning and voice 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 [241] 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 [159-161, 196]. 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 hardly generalizes well for other types of spoofing attacks [242]. Moreover, there does

not exist a unified countermeasure that can detect replay and cloning attacks in multi-hop scenarios, where multiple microphones and smart speakers are chained together. We addressed the problem of spoofing attack detection on multi-hop scenarios in our prior work [10], 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.

- Integrated ASV with anti-spoofing: The exponential growth of smart speakers and other voice-enabled devices considers 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.
- Chained cloned and cloned replay attack detection in smart speakers. 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) [9,10].

8 Conclusion

This survey paper presents a comprehensive review of existing deepfake generation and detection methods. Not all digital manipulations are harmful. However, due to immense technological advancements it is now very easy to produce realistic fabricated content. Therefore, malicious users can use it to spread disinformation to attack individuals and 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.

References

- [1] (11.09.2020). ZAO. Available: <https://apps.apple.com/cn/app/zao/id1465199127>.
- [2] (11.09.2020). Reface App. Available: <https://reface.app/>
- [3] (17.09.2020). FaceApp. Available: <https://www.faceapp.com/>
- [4] (07.09.2020). Audacity. Available: <https://www.audacityteam.org/>
- [5] (11.01.2021). Sound Forge. Available: <https://www.magix.com/gb/music/sound-forge/>
- [6] J. F. Boylan, "Will deep-fake technology destroy democracy?", *The New York Times*, Oct, vol. 17, 2018.
- [7] D. Harwell, "Scarlett Johansson on fake AI-generated sex videos: 'Nothing can stop someone from cutting and pasting my image,'" *Washington Post*, 2018.
- [8] C. Chan, S. Ginosar, T. Zhou, and A. A. Efros, "Everybody Dance Now," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 5933-5942.
- [9] K. M. Malik, H. Malik, and R. Baumann, "Towards vulnerability analysis of voice-driven interfaces and countermeasures for replay attacks," in *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, 2019, pp. 523-528: IEEE.
- [10] K. M. Malik, A. Javed, H. Malik, and A. Irtaza, "A light-weight replay detection framework for voice controlled iot devices," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 5, pp. 982-996, 2020.
- [11] D. Harwell. (2019). *An artificial-intelligence first: Voice-mimicking software reportedly used in a major theft*. Available: <https://www.washingtonpost.com/technology/2019/09/04/an-artificial-intelligence-first-voice-mimicking-software-reportedly-used-major-theft/>
- [12] L. Verdoliva, "Media forensics and deepfakes: an overview," *arXiv preprint arXiv:2001.06564*, 2020.
- [13] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "Deepfakes and beyond: A survey of face manipulation and fake detection," *arXiv preprint arXiv:2001.00179*, 2020.

- [14] T. T. Nguyen, C. M. Nguyen, D. T. Nguyen, D. T. Nguyen, and S. Nahavandi, "Deep Learning for Deepfakes Creation and Detection," *arXiv preprint arXiv:1909.11573*, 2019.
- [15] Y. Mirsky and W. Lee, "The Creation and Detection of Deepfakes: A Survey," *arXiv preprint arXiv:2004.11138*, 2020.
- [16] L. Oliveira, "The current state of fake news," *Procedia Computer Science*, vol. 121, no. C, pp. 817-825, 2017.
- [17] R. Chesney and D. Citron, "Deepfakes and the New Disinformation War: The Coming Age of Post-Truth Geopolitics," *Foreign Aff.*, vol. 98, p. 147, 2019.
- [18] W. Phillips, *This is why we can't have nice things: Mapping the relationship between online trolling and mainstream culture*. MIT Press, 2015.
- [19] T. Higgin, "FCJ-159/b/lack up: What trolls can teach us about race," *The Fibreculture Journal*, no. 22 2013: Trolls and The Negative Space of the Internet, 2013.
- [20] T. Mihaylov, G. Georgiev, and P. Nakov, "Finding opinion manipulation trolls in news community forums," in *Proceedings of the nineteenth conference on computational natural language learning*, 2015, pp. 310-314.
- [21] T. P. Gerber and J. Zavisca, "Does Russian propaganda work?," *The Washington Quarterly*, vol. 39, no. 2, pp. 79-98, 2016.
- [22] P. N. Howard and B. Kollanyi, "Bots, StrongerIn, and Brexit: computational propaganda during the UK-EU referendum," 2016, Art. no. Available at SSRN 2798311.
- [23] O. Varol, E. Ferrara, C. A. Davis, F. Menczer, and A. Flammini, "Online human-bot interactions: Detection, estimation, and characterization," in *Eleventh international AAAI conference on web and social media*, 2017.
- [24] H. Setiaji and I. V. Paputungan, "Design of telegram bots for campus information sharing," in *IOP Conference Series: Materials Science and Engineering*, 2018, vol. 325, no. 1, p. 012005: Institute of Physics Publishing.
- [25] A. Marwick and R. Lewis, "Media manipulation and disinformation online," *New York: Data Society Research Institute*, 2017.
- [26] C. R. Sunstein and A. Vermeule, "Conspiracy theories: Causes and cures," *Journal of Political Philosophy*, vol. 17, no. 2, pp. 202-227, 2009.
- [27] R. Faris, H. Roberts, B. Etling, N. Bourassa, E. Zuckerman, and Y. Benkler, "Partisanship, propaganda, and disinformation: Online media and the 2016 US presidential election," *Berkman Klein Center Research Publication*, vol. 6, 2017.
- [28] L. Benedictus, "Invasion of the troll armies: from Russian Trump supporters to Turkish state stooges," *The Guardian*, vol. 6, p. 2016, 2016.
- [29] N. A. Mhiripiri and T. Chari, *Media law, ethics, and policy in the digital age*. IGI Global, 2017.
- [30] H. Huang, P. S. Yu, and C. Wang, "An introduction to image synthesis with generative adversarial nets," *arXiv preprint arXiv:1803.04469*, 2018.
- [31] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8789-8797.
- [32] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman, "Synthesizing Obama: learning lip sync from audio," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 95:1-95:13, 2017.
- [33] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2face: Real-time face capture and reenactment of rgb videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2387-2395.
- [34] O. Wiles, A. Sophia Koepke, and A. Zisserman, "X2face: A network for controlling face generation using images, audio, and pose codes," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 670-686.

- [35] B. Paris and J. Donovan, "Deepfakes and Cheap Fakes," *United States of America: Data & Society*, 2019.
- [36] C. Bregler, M. Covell, and M. Slaney, "Video rewrite: Driving visual speech with audio," in *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, 1997, pp. 353-360.
- [37] J. Vincent. (18.09.2020). *New AI deepfake app creates nude images of women*. Available: <https://www.theverge.com/2019/6/27/18760896/deepfake-nude-ai-app-women-deepnude-non-consensual-pornography>
- [38] (18.09.2020). *FakeApp* 2.2.0. Available: <https://www.malavida.com/en/soft/fakeapp/>
- [39] (18.09.2020). *Faceswap: Deepfakes software for all*. Available: <https://github.com/deepfakes/faceswap>
- [40] (18.09.2020). *DeepFaceLab*. Available: <https://github.com/iperov/DeepFaceLab>
- [41] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, "First order motion model for image animation," in *Advances in Neural Information Processing Systems*, 2019, pp. 7137-7147.
- [42] H. Kim *et al.*, "Deep video portraits," *ACM Trans. Graph.*, vol. 37, no. 4, pp. 163:1-163:14, 2018.
- [43] S. Ha, M. Kersner, B. Kim, S. Seo, and D. Kim, "Marionette: Few-shot face reenactment preserving identity of unseen targets," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, no. 07, pp. 10893-10900.
- [44] Y. Wang, P. Bilinski, F. Bremond, and A. Dantcheva, "ImaGINator: Conditional Spatio-Temporal GAN for Video Generation," in *The IEEE Winter Conference on Applications of Computer Vision*, 2020, pp. 1160-1169.
- [45] M. Westerlund, "The emergence of deepfake technology: A review," *Technology Innovation Management Review*, vol. 9, no. 11, 2019.
- [46] S. Greengard, "Will deepfakes do deep damage?," ed: ACM New York, NY, USA, 2019.
- [47] A. v. d. Oord *et al.*, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.
- [48] Y. Wang *et al.*, "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017.
- [49] S. O. Arik *et al.*, "Deep voice: Real-time neural text-to-speech," *arXiv preprint arXiv:1702.07825*, 2017.
- [50] S. Arik, J. Chen, K. Peng, W. Ping, and Y. Zhou, "Neural voice cloning with a few samples," in *Advances in Neural Information Processing Systems*, 2018, pp. 10019-10029.
- [51] I. Goodfellow *et al.*, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672-2680.
- [52] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th international conference on Machine learning*, 2008, pp. 1096-1103.
- [53] C. Donahue, J. McAuley, and M. Puckette, "Synthesizing audio with generative adversarial networks," *arXiv preprint arXiv:1802.04208*, 2018.
- [54] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," *arXiv preprint arXiv:1605.05396*, 2016.
- [55] D. Nie *et al.*, "Medical image synthesis with context-aware generative adversarial networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2017, pp. 417-425: Springer.
- [56] Y. Nirkin, I. Masi, A. T. Tuan, T. Hassner, and G. Medioni, "On face segmentation, face swapping, and face perception," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 2018, pp. 98-105: IEEE.

- [57] J. Thies, M. Zollhöfer, C. Theobalt, M. Stamminger, and M. Nießner, "real-time reenactment of human portrait videos," *ACM Trans. Graph.*, vol. 37, no. 4, pp. 164:1-164:13, / 2018.
- [58] B. Dolhansky *et al.*, "The DeepFake Detection Challenge Dataset," *arXiv preprint arXiv:2006.07397*, 2020.
- [59] N. Neverova, R. Alp Guler, and I. Kokkinos, "Dense pose transfer," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 123-138.
- [60] D. Bitouk, N. Kumar, S. Dhillon, P. Belhumeur, and S. K. Nayar, "Face swapping: automatically replacing faces in photographs," in *ACM Transactions on Graphics (TOG)*, 2008, vol. 27, no. 3, p. 39: ACM.
- [61] Y. Lin, Q. Lin, F. Tang, and S. Wang, "Face replacement with large-pose differences," in *Proceedings of the 20th ACM international conference on Multimedia*, 2012, pp. 1249-1250: ACM.
- [62] B. M. Smith and L. Zhang, "Joint face alignment with non-parametric shape models," in *European Conference on Computer Vision*, 2012, pp. 43-56: Springer.
- [63] (18.09.2020). *DFaker*. Available: <https://github.com/dfaker/df>
- [64] (18.09.2020). *DeepFake-tf: Deepfake based on tensorflow*. Available: <https://github.com/StromWine/DeepFake-tf>
- [65] (18.09.2020). *Faceswap-GAN* Available: <https://github.com/shaoanlu/faceswap-GAN>
- [66] I. Korshunova, W. Shi, J. Dambre, and L. Theis, "Fast face-swap using convolutional neural networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3677-3685.
- [67] Y. Nirkin, Y. Keller, and T. Hassner, "FSGAN: Subject Agnostic Face Swapping and Reenactment," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 7184-7193.
- [68] R. Natsume, T. Yatagawa, and S. Morishima, "Rsgan: face swapping and editing using face and hair representation in latent spaces," *arXiv preprint arXiv:1804.03447*, 2018.
- [69] R. Natsume, T. Yatagawa, and S. Morishima, "Fsnet: An identity-aware generative model for image-based face swapping," in *Asian Conference on Computer Vision*, 2018, pp. 117-132: Springer.
- [70] L. Li, J. Bao, H. Yang, D. Chen, and F. Wen, "Faceshifter: Towards high fidelity and occlusion aware face swapping," *arXiv preprint arXiv:1912.13457*, 2019.
- [71] B. Fan, L. Wang, F. K. Soong, and L. Xie, "Photo-real talking head with deep bidirectional LSTM," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4884-4888: IEEE.
- [72] J. Charles, D. Magee, and D. Hogg, "Virtual immortality: Reanimating characters from tv shows," in *European Conference on Computer Vision*, 2016, pp. 879-886: Springer.
- [73] A. Jamaludin, J. S. Chung, and A. Zisserman, "You said that?: Synthesising talking faces from audio," *International Journal of Computer Vision*, pp. 1-13, 2019.
- [74] K. Vougioukas, S. Petridis, and M. Pantic, "End-to-End Speech-Driven Realistic Facial Animation with Temporal GANs," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 37-40.
- [75] H. Zhou, Y. Liu, Z. Liu, P. Luo, and X. Wang, "Talking face generation by adversarially disentangled audio-visual representation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, pp. 9299-9306.
- [76] P. Garrido *et al.*, "Vdub: Modifying face video of actors for plausible visual alignment to a dubbed audio track," in *Computer graphics forum*, 2015, vol. 34, no. 2, pp. 193-204: Wiley Online Library.

- [77] P. KR, R. Mukhopadhyay, J. Philip, A. Jha, V. Namboodiri, and C. Jawahar, "Towards automatic face-to-face translation," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 1428-1436.
- [78] K. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. Jawahar, "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*, 2020, pp. 484-492.
- [79] O. Fried *et al.*, "Text-based editing of talking-head video," *ACM Transactions on Graphics (TOG)*, vol. 38, no. 4, pp. 1-14, 2019.
- [80] B.-H. Kim and V. Ganapathi, "LumiereNet: Lecture Video Synthesis from Audio," *arXiv preprint arXiv:1907.02253*, 2019.
- [81] J. Thies, M. Zollhöfer, M. Nießner, L. Valgaerts, M. Stamminger, and C. Theobalt, "Real-time expression transfer for facial reenactment," *ACM Trans. Graph.*, vol. 34, no. 6, pp. 183:1-183:14, 2015.
- [82] M. Zollhöfer *et al.*, "Real-time non-rigid reconstruction using an RGB-D camera," *ACM Transactions on Graphics (ToG)*, vol. 33, no. 4, pp. 1-12, 2014.
- [83] J. Thies, M. Zollhöfer, and M. Nießner, "IMU2Face: Real-time Gesture-driven Facial Reenactment," *arXiv preprint arXiv:1801.01446*, 2017.
- [84] J. Thies, M. Zollhöfer, C. Theobalt, M. Stamminger, and M. Nießner, "Headon: Real-time reenactment of human portrait videos," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 4, pp. 1-13, 2018.
- [85] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8798-8807.
- [86] H. Kim *et al.*, "Deep video portraits," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 4, p. 163, 2018.
- [87] W. Wu, Y. Zhang, C. Li, C. Qian, and C. Change Loy, "Reenactgan: Learning to reenact faces via boundary transfer," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 603-619.
- [88] A. Pumarola, A. Agudo, A. M. Martínez, A. Sanfeliu, and F. Moreno-Noguer, "GANimation: Anatomically-Aware Facial Animation from a Single Image," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 818-833.
- [89] E. Sanchez and M. Valstar, "Triple consistency loss for pairing distributions in GAN-based face synthesis," *arXiv preprint arXiv:1811.03492*, 2018.
- [90] E. Zakharov, A. Shysheya, E. Burkov, and V. Lempitsky, "Few-shot adversarial learning of realistic neural talking head models," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 9459-9468.
- [91] Y. Zhang, S. Zhang, Y. He, C. Li, C. C. Loy, and Z. Liu, "One-shot face reenactment," *arXiv preprint arXiv:1908.03251*, 2019.
- [92] H. Hao, S. Baireddy, A. R. Reibman, and E. J. Delp, "FaR-GAN for One-Shot Face Reenactment," *arXiv preprint arXiv:2005.06402*, 2020.
- [93] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces," in *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, 1999, pp. 187-194.
- [94] J. Lorenzo-Trueba *et al.*, "The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods," *arXiv preprint arXiv:1804.04262*, 2018.
- [95] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.
- [96] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

- [97] M.-Y. Liu and O. Tuzel, "Coupled generative adversarial networks," in *Advances in neural information processing systems*, 2016, pp. 469-477.
- [98] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.
- [99] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 4401-4410.
- [100] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8110-8119.
- [101] R. Huang, S. Zhang, T. Li, and R. He, "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*, 2017, pp. 2439-2448.
- [102] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *International Conference on Machine Learning*, 2019, pp. 7354-7363: PMLR.
- [103] A. Brock, J. Donahue, and K. Simonyan, "Large scale gan training for high fidelity natural image synthesis," *arXiv preprint arXiv:1809.11096*, 2018.
- [104] H. Zhang *et al.*, "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5907-5915.
- [105] G. Perarnau, J. Van De Weijer, B. Raducanu, and J. M. Álvarez, "Invertible conditional gans for image editing," *arXiv preprint arXiv:1611.06355*, 2016.
- [106] G. Lample, N. Zeghidour, N. Usunier, A. Bordes, L. Denoyer, and M. A. Ranzato, "Fader networks: Manipulating images by sliding attributes," in *Advances in neural information processing systems*, 2017, pp. 5967-5976.
- [107] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha, "Stargan v2: Diverse image synthesis for multiple domains," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8188-8197.
- [108] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen, "Attgan: Facial attribute editing by only changing what you want," *IEEE Transactions on Image Processing*, vol. 28, no. 11, pp. 5464-5478, 2019.
- [109] M. Liu *et al.*, "Stgan: A unified selective transfer network for arbitrary image attribute editing," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 3673-3682.
- [110] I. Petrov *et al.*, "DeepFaceLab: A simple, flexible and extensible face swapping framework," *arXiv preprint arXiv:2005.05535*, 2020.
- [111] D. Chen, Q. Chen, J. Wu, X. Yu, and T. Jia, "Face Swapping: Realistic Image Synthesis Based on Facial Landmarks Alignment," *Mathematical Problems in Engineering*, vol. 2019, 2019.
- [112] G. Zhang, M. Kan, S. Shan, and X. Chen, "Generative adversarial network with spatial attention for face attribute editing," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 417-432.
- [113] Z. He, M. Kan, J. Zhang, and S. Shan, "PA-GAN: Progressive Attention Generative Adversarial Network for Facial Attribute Editing," *arXiv preprint arXiv:2007.05892*, 2020.
- [114] J. Lorenzo-Trueba, F. Fang, X. Wang, I. Echizen, J. Yamagishi, and T. Kinnunen, "Can we steal your vocal identity from the Internet?: Initial investigation of cloning Obama's voice using GAN, WaveNet and low-quality found data," *arXiv preprint arXiv:1803.00860*, 2018.
- [115] Z. Jin, G. J. Mysore, S. Diverdi, J. Lu, and A. Finkelstein, "Voco: Text-based insertion and replacement in audio narration," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, pp. 1-13, 2017.

- [116] J. Sotelo *et al.*, "Char2wav: End-to-end speech synthesis," 2017.
- [117] A. Van den Oord, N. Kalchbrenner, L. Espeholt, O. Vinyals, and A. Graves, "Conditional image generation with pixelcnn decoders," in *Advances in neural information processing systems*, 2016, pp. 4790-4798.
- [118] M. Zhang, B. Sisman, L. Zhao, and H. Li, "DeepConversion: Voice conversion with limited parallel training data," *Speech Communication*, 2020.
- [119] P. Partila, J. Tovarek, G. H. Ilk, J. Rozhon, and M. Voznak, "Deep Learning Serves Voice Cloning: How Vulnerable Are Automatic Speaker Verification Systems to Spoofing Trials?," *IEEE Communications Magazine*, vol. 58, no. 2, pp. 100-105, 2020.
- [120] W. Ping *et al.*, "Deep voice 3: 2000-speaker neural text-to-speech," *Proc. ICLR*, pp. 214-217, 2018.
- [121] A. Oord *et al.*, "Parallel wavenet: Fast high-fidelity speech synthesis," in *International conference on machine learning*, 2018, pp. 3918-3926: PMLR.
- [122] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236-243, 1984.
- [123] A. Gibiansky *et al.*, "Deep voice 2: Multi-speaker neural text-to-speech," in *Advances in neural information processing systems*, 2017, pp. 2962-2970.
- [124] Y. Yasuda, X. Wang, S. Takaki, and J. Yamagishi, "Investigation of enhanced Tacotron text-to-speech synthesis systems with self-attention for pitch accent language," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6905-6909: IEEE.
- [125] Y. Taigman, L. Wolf, A. Polyak, and E. Nachmani, "Voiceloop: Voice fitting and synthesis via a phonological loop," *arXiv preprint arXiv:1707.06588*, 2017.
- [126] Y. Jia *et al.*, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in *Advances in neural information processing systems*, 2018, pp. 4480-4490.
- [127] H.-T. Luong and J. Yamagishi, "NAUTILUS: a Versatile Voice Cloning System," *arXiv preprint arXiv:2005.11004*, 2020.
- [128] Y. Chen *et al.*, "Sample efficient adaptive text-to-speech," *arXiv preprint arXiv:1809.10460*, 2018.
- [129] J. Cong, S. Yang, L. Xie, G. Yu, and G. Wan, "Data Efficient Voice Cloning from Noisy Samples with Domain Adversarial Training," *arXiv preprint arXiv:2008.04265*, 2020.
- [130] K. Nagano *et al.*, "paGAN: real-time avatars using dynamic textures," *ACM Trans. Graph.*, vol. 37, no. 6, pp. 258:1-258:12, 2018.
- [131] A. Pumarola, A. Agudo, A. M. Martinez, A. Sanfeliu, and F. Moreno-Noguer, "Ganimation: Anatomically-aware facial animation from a single image," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 818-833.
- [132] J. Thies, M. Zollhöfer, C. Theobalt, M. Stamminger, and M. Nießner, "Headon: Real-time reenactment of human portrait videos," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 4, p. 164, 2018.
- [133] P. Korshunov and S. Marcel, "Speaker inconsistency detection in tampered video," in *2018 26th European Signal Processing Conference (EUSIPCO)*, 2018, pp. 2375-2379: IEEE.
- [134] F. Matern, C. Riess, and M. Stamminger, "Exploiting visual artifacts to expose deepfakes and face manipulations," in *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, 2019, pp. 83-92: IEEE.
- [135] T. Jung, S. Kim, and K. Kim, "DeepVision: Deepfakes Detection Using Human Eye Blinking Pattern," *IEEE Access*, vol. 8, pp. 83144-83154, 2020.
- [136] R. Durall, M. Keuper, F.-J. Pfreundt, and J. Keuper, "Unmasking deepfakes with simple features," *arXiv preprint arXiv:00686*, 2019.

- [137] T. Zhao, X. Xu, M. Xu, H. Ding, Y. Xiong, and W. Xia, "Learning to Recognize Patch-Wise Consistency for Deepfake Detection," *arXiv preprint arXiv:09311*, 2020.
- [138] E. A. AlBadawy, S. Lyu, and H. Farid, "Detecting AI-Synthesized Speech Using Bispectral Analysis," in *CVPR Workshops*, 2019, pp. 104-109.
- [139] H. Malik, "Securing voice-driven interfaces against fake (cloned) audio attacks," in *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, 2019, pp. 512-517: IEEE.
- [140] S. McCloskey and M. Albright, "Detecting gan-generated imagery using color cues," *arXiv preprint arXiv:08247*, 2018.
- [141] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-df: A new dataset for deepfake forensics," *arXiv preprint arXiv:1909.12962*, 2019.
- [142] X. Li, K. Yu, S. Ji, Y. Wang, C. Wu, and H. Xue, "Fighting against deepfake: Patch & pair convolutional neural networks (ppcnn)," in *Companion Proceedings of the Web Conference 2020*, 2020, pp. 88-89.
- [143] X. Yang, Y. Li, and S. Lyu, "Exposing deep fakes using inconsistent head poses," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 8261-8265: IEEE.
- [144] N. Yu, L. S. Davis, and M. Fritz, "Attributing fake images to GANs: Learning and analyzing GAN fingerprints," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 7556-7566.
- [145] Z. Guo, G. Yang, J. Chen, and X. Sun, "Fake face detection via adaptive residuals extraction network," *arXiv preprint arXiv:04945*, 2020.
- [146] S. Agarwal, H. Farid, Y. Gu, M. He, K. Nagano, and H. Li, "Protecting World Leaders Against Deep Fakes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 38-45.
- [147] S. Agarwal, T. El-Gaaly, H. Farid, and S.-N. Lim, "Detecting Deep-Fake Videos from Appearance and Behavior," *arXiv preprint arXiv:14491*, 2020.
- [148] U. A. Ciftci and I. Demir, "FakeCatcher: Detection of Synthetic Portrait Videos using Biological Signals," *arXiv preprint arXiv:1901.02212*, 2019.
- [149] Y. Li, M.-C. Chang, and S. Lyu, "In ictu oculi: Exposing ai generated fake face videos by detecting eye blinking," *arXiv preprint arXiv:1806.02877*, 2018.
- [150] S. Fernandes *et al.*, "Predicting Heart Rate Variations of Deepfake Videos using Neural ODE," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019, pp. 0-0.
- [151] I. Amerini and R. Caldelli, "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*, 2020, pp. 97-102.
- [152] J. Fei, Z. Xia, P. Yu, and F. Xiao, "Exposing AI-generated videos with motion magnification," *Multimedia Tools Applications*, pp. 1-14, 2020.
- [153] D. Güera, S. Baireddy, P. Bestagini, S. Tubaro, and E. J. Delp, "We Need No Pixels: Video Manipulation Detection Using Stream Descriptors," *arXiv preprint arXiv:1906.08743*, 2019.
- [154] D. Güera and E. J. Delp, "Deepfake video detection using recurrent neural networks," in *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2018, pp. 1-6: IEEE.
- [155] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan, "Recurrent Convolutional Strategies for Face Manipulation Detection in Videos," *Interfaces (GUI)*, vol. 3, p. 1, 2019.
- [156] I. Amerini, L. Galteri, R. Caldelli, and A. Del Bimbo, "Deepfake Video Detection through Optical Flow based CNN," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019, pp. 0-0.

- [157] Y. Zhang, L. Zheng, and V. L. Thing, "Automated face swapping and its detection," in *2017 IEEE 2nd International Conference on Signal and Image Processing (ICSIP)*, 2017, pp. 15-19: IEEE.
- [158] P. Korshunov and S. Marcel, "Deepfakes: a new threat to face recognition? assessment and detection," *arXiv preprint arXiv:1812.08685*, 2018.
- [159] L. Guarnera, O. Giudice, and S. Battiato, "DeepFake Detection by Analyzing Convolutional Traces," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 666-667.
- [160] Y. Li and S. Lyu, "Exposing deepfake videos by detecting face warping artifacts," *arXiv preprint arXiv:1811.00656*, vol. 2, 2018.
- [161] L. Nataraj *et al.*, "Detecting GAN generated fake images using co-occurrence matrices," *arXiv preprint arXiv:1903.06836*, 2019.
- [162] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics++: Learning to detect manipulated facial images," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 1-11.
- [163] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "Mesonet: a compact facial video forgery detection network," in *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, 2018, pp. 1-7: IEEE.
- [164] H. H. Nguyen, F. Fang, J. Yamagishi, and I. Echizen, "Multi-task learning for detecting and segmenting manipulated facial images and videos," *arXiv preprint arXiv:1906.06876*, 2019.
- [165] D. Cozzolino, J. Thies, A. Rössler, C. Riess, M. Nießner, and L. Verdoliva, "Forensictransfer: Weakly-supervised domain adaptation for forgery detection," *arXiv preprint arXiv:1812.02510*, 2018.
- [166] D. M. Montserrat *et al.*, "Deepfakes Detection with Automatic Face Weighting," *arXiv preprint arXiv:12027*, 2020.
- [167] O. de Lima, S. Franklin, S. Basu, B. Karwoski, and A. George, "Deepfake Detection using Spatiotemporal Convolutional Networks," *arXiv preprint arXiv:14749*, 2020.
- [168] F. Marra, C. Saltori, G. Boato, and L. Verdoliva, "Incremental learning for the detection and classification of GAN-generated images," in *2019 IEEE International Workshop on Information Forensics and Security (WIFS)*, 2019, pp. 1-6: IEEE.
- [169] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "CNN-generated images are surprisingly easy to spot... for now," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8695-8704.
- [170] F. Marra, D. Gragnaniello, D. Cozzolino, and L. Verdoliva, "Detection of gan-generated fake images over social networks," in *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, 2018, pp. 384-389: IEEE.
- [171] A. Chintha *et al.*, "Recurrent convolutional structures for audio spoof and video deepfake detection," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 5, pp. 1024-1037, 2020.
- [172] S. Agarwal, H. Farid, O. Fried, and M. Agrawala, "Detecting deep-fake videos from phoneme-viseme mismatches," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 660-661.
- [173] M. S. Rana and A. H. Sung, "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)*, 2020, pp. 70-75: IEEE.
- [174] R. Wang *et al.*, "DeepSonar: Towards Effective and Robust Detection of AI-Synthesized Fake Voices," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 1207-1216.

- [175] A. K. Singh and P. Singh, "Detection of AI-Synthesized Speech Using Cepstral & Bispectral Statistics," *arXiv preprint arXiv:2009.01934*, 2020.
- [176] H. Malik, "Fighting AI with AI: Fake Speech Detection Using Deep Learning," in *2019 AES International Conference on Audio Forensics*, 2019.
- [177] Z. Wu, R. K. Das, J. Yang, and H. Li, "Light Convolutional Neural Network with Feature Genuinization for Detection of Synthetic Speech Attacks," *arXiv preprint arXiv:09637*, 2020.
- [178] R. Wang, L. Ma, F. Juefei-Xu, X. Xie, J. Wang, and Y. Liu, "Fakespotter: A simple baseline for spotting ai-synthesized fake faces," *arXiv preprint arXiv:06122*, 2019.
- [179] H. Khalid and S. S. Woo, "OC-FakeDect: Classifying deepfakes using one-class variational autoencoder," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 656-657.
- [180] K. Chugh, P. Gupta, A. Dhall, and R. Subramanian, "Not made for each other-Audio-Visual Dissonance-based Deepfake Detection and Localization," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 439-447.
- [181] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "Emotions Don't Lie: An Audio-Visual Deepfake Detection Method using Affective Cues," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 2823-2832.
- [182] D. Cozzolino, A. Rössler, J. Thies, M. Nießner, and L. Verdoliva, "ID-Reveal: Identity-aware DeepFake Video Detection," *arXiv preprint arXiv:02512*, 2020.
- [183] J.-L. Zhong, C.-M. Pun, and Y.-F. Gan, "Dense Moment Feature Index and Best Match Algorithms for Video Copy-Move Forgery Detection," *Information Sciences*, 2020.
- [184] X. Ding, Y. Huang, Y. Li, and J. He, "Forgery detection of motion compensation interpolated frames based on discontinuity of optical flow," *Multimedia Tools Applications*, pp. 1-26, 2020.
- [185] P. Niyishaka and C. Bhagvati, "Copy-move forgery detection using image blobs and BRISK feature," *Multimedia Tools Applications*, pp. 1-15, 2020.
- [186] M. Abdel-Basset, G. Manogaran, A. E. Fakhry, and I. El-Henawy, "2-Levels of clustering strategy to detect and locate copy-move forgery in digital images," *Multimedia Tools Applications*, vol. 79, no. 7, pp. 5419-5437, 2020.
- [187] Z. Akhtar and D. Dasgupta, "A comparative evaluation of local feature descriptors for deepfakes detection," in *2019 IEEE International Symposium on Technologies for Homeland Security (HST)*, 2019, pp. 1-5: IEEE.
- [188] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815-823.
- [189] A. J. O'Toole, P. J. Phillips, F. Jiang, J. Ayyad, N. Penard, and H. Abdi, "Face recognition algorithms surpass humans matching faces over changes in illumination," *IEEE transactions on pattern analysis machine intelligence*, vol. 29, no. 9, pp. 1642-1646, 2007.
- [190] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Openface: an open source facial behavior analysis toolkit," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2016, pp. 1-10: IEEE.
- [191] K. Jack, "Chapter 13-MPEG-2," *Video Demystified: A Handbook for the Digital Engineer*, pp. 577-737.
- [192] C. Sanderson, "The vidtimit database," IDIAP2002.
- [193] A. Anand, R. D. Labati, A. Genovese, E. Muñoz, V. Piuri, and F. Scotti, "Age estimation based on face images and pre-trained convolutional neural networks," in *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2017, pp. 1-7: IEEE.

- [194] E. Boutellaa, Z. Boulkenafet, J. Komulainen, and A. Hadid, "Audiovisual synchrony assessment for replay attack detection in talking face biometrics," *Multimedia Tools Applications*, vol. 75, no. 9, pp. 5329-5343, 2016.
- [195] R. Ranjan, V. M. Patel, and R. Chellappa, "Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition," *IEEE Transactions on Pattern Analysis Machine Intelligence*, vol. 41, no. 1, pp. 121-135, 2017.
- [196] T. Soukupova and J. Cech, "Eye blink detection using facial landmarks," in *21st computer vision winter workshop, Rimske Toplice, Slovenia*, 2016.
- [197] K. I. Laws, "Textured image segmentation," University of Southern California Los Angeles Image Processing INST1980.
- [198] P. Nagarsheth, E. Khoury, K. Patil, and M. Garland, "Replay Attack Detection Using DNN for Channel Discrimination," in *Interspeech*, 2017, pp. 97-101.
- [199] T. Gunendradasan, S. Irtza, E. Ambikairajah, and J. Epps, "Transmission line cochlear model based am-fm features for replay attack detection," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6136-6140: IEEE.
- [200] M. Witkowski, S. Kacprzak, P. Zelasko, K. Kowalczyk, and J. Galka, "Audio Replay Attack Detection Using High-Frequency Features," in *Interspeech*, 2017, pp. 27-31.
- [201] M. Saranya, R. Padmanabhan, and H. A. Murthy, "Replay attack detection in speaker verification using non-voiced segments and decision level feature switching," in *2018 International Conference on Signal Processing and Communications (SPCOM)*, 2018, pp. 332-336: IEEE.
- [202] D. E. King, "Dlib-ml: A machine learning toolkit," *The Journal of Machine Learning Research*, vol. 10, pp. 1755-1758, 2009.
- [203] L. Alparone, M. Barni, F. Bartolini, and R. Caldelli, "Regularization of optic flow estimates by means of weighted vector median filtering," *IEEE transactions on image processing*, vol. 8, no. 10, pp. 1462-1467, 1999.
- [204] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8934-8943.
- [205] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499-1503, 2016.
- [206] O. Wiles, A. Koepke, and A. Zisserman, "Self-supervised learning of a facial attribute embedding from video," *arXiv preprint arXiv:06882*, 2018.
- [207] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," *arXiv preprint arXiv:* 2014.
- [208] H. Rahman, M. U. Ahmed, S. Begum, and P. Funk, "Real time heart rate monitoring from facial RGB color video using webcam," in *The 29th Annual Workshop of the Swedish Artificial Intelligence Society (SAIS), 2-3 June 2016, Malmö, Sweden*, 2016, no. 129: Linköping University Electronic Press.
- [209] H.-Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, and W. Freeman, "Eulerian video magnification for revealing subtle changes in the world," *ACM transactions on graphics*, vol. 31, no. 4, pp. 1-8, 2012.
- [210] R. T. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud, "Neural ordinary differential equations," in *Advances in neural information processing systems*, 2018, pp. 6571-6583.
- [211] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," 2015.
- [212] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert, "ICARL: Incremental Classifier and Representation Learning," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 2001-2010.

- [213] T. Chen, A. Kumar, P. Nagarsheth, G. Sivaraman, and E. Khoury, "Generalization of Audio Deepfake Detection," in *Proc. Odyssey 2020 The Speaker and Language Recognition Workshop*, 2020, pp. 132-137.
- [214] L. Huang and C.-M. Pun, "Audio Replay Spoof Attack Detection by Joint Segment-Based Linear Filter Bank Feature Extraction and Attention-Enhanced DenseNet-BiLSTM Network," *IEEE/ACM Transactions on Audio, Speech, Language Processing*, vol. 28, pp. 1813-1825, 2020.
- [215] C.-I. Lai, A. Abad, K. Richmond, J. Yamagishi, N. Dehak, and S. King, "Attentive filtering networks for audio replay attack detection," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6316-6320: IEEE.
- [216] J. Yang and R. K. Das, "Long-term high frequency features for synthetic speech detection," *Digital Signal Processing*, vol. 97, p. 102622, 2020.
- [217] M. Todisco *et al.*, "Asvspoof 2019: Future horizons in spoofed and fake audio detection," *arXiv preprint arXiv:05441*, 2019.
- [218] P. Korshunov *et al.*, "Overview of BTAS 2016 speaker anti-spoofing competition," in *2016 IEEE 8th international conference on biometrics theory, applications and systems (BTAS)*, 2016, pp. 1-6: IEEE.
- [219] H. Delgado *et al.*, "ASVspoof 2017 Version 2.0: meta-data analysis and baseline enhancements," 2018.
- [220] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics: A large-scale video dataset for forgery detection in human faces," *arXiv preprint arXiv:1803.09179*, 2018.
- [221] (14.08.2020). *Faceswap*. Available: <https://github.com/MarekKowalski/FaceSwap/>
- [222] J. Thies, M. Zollhöfer, and M. Nießner, "Deferred neural rendering: Image synthesis using neural textures," *ACM Transactions on Graphics (TOG)*, vol. 38, no. 4, pp. 1-12, 2019.
- [223] S. Abu-El-Haija *et al.*, "Youtube-8m: A large-scale video classification benchmark," *arXiv preprint arXiv:1609.08675*, 2016.
- [224] A. Aravkin, J. V. Burke, L. Ljung, A. Lozano, and G. Pillonetto, "Generalized Kalman smoothing: Modeling and algorithms," *Automatica*, vol. 86, pp. 63-86, 2017.
- [225] E. Reinhard, M. Adhikhmin, B. Gooch, and P. Shirley, "Color transfer between images," *IEEE Computer graphics*, vol. 21, no. 5, pp. 34-41, 2001.
- [226] K. Ito. (25.12.2020). *The LJ speech dataset*. Available: <https://keithito.com/LJ-Speech-Dataset>
- [227] M. A. I. L. GmbH. (25.02.2021). *The MAILABS speech dataset*. Available: <https://www.caito.de/2019/01/the-m-ailabs-speech-dataset/>
- [228] R. Ardila *et al.*, "Common voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670*, 2019.
- [229] X. Wang *et al.*, "ASVspoof 2019: a large-scale public database of synthetized, converted and replayed speech," *Computer Speech & Language*, p. 101114, 2020.
- [230] R. Reimao and V. Tzerpos, "FoR: A Dataset for Synthetic Speech Detection," in *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, 2019, pp. 1-10: IEEE.
- [231] A. Bansal, S. Ma, D. Ramanan, and Y. Sheikh, "Recycle-gan: Unsupervised video retargeting," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 119-135.
- [232] R. Natsume, T. Yatagawa, and S. Morishima, "FSNet: An Identity-Aware Generative Model for Image-Based Face Swapping," presented at the Asian Conference on Computer Vision, Cham, 2018.
- [233] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," *Journal of the Acoustical Society of Japan (E)*, vol. 11, no. 2, pp. 71-76, 1990.

- [234] P. Fraga-Lamas and T. M. Fernández-Caramés, "Fake News, Disinformation, and Deepfakes: Leveraging Distributed Ledger Technologies and Blockchain to Combat Digital Deception and Counterfeit Reality," *IT Professional*, vol. 22, no. 2, pp. 53-59, 2020.
- [235] H. R. Hasan and K. Salah, "Combating deepfake videos using blockchain and smart contracts," *IEEE Access*, vol. 7, pp. 41596-41606, 2019.
- [236] M. Feng and H. Xu, "Deep reinforcement learning based optimal defense for cyber-physical system in presence of unknown cyber-attack," in *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2017, pp. 1-8: IEEE.
- [237] X. Xu and T. Xie, "A reinforcement learning approach for host-based intrusion detection using sequences of system calls," in *International Conference on Intelligent Computing*, 2005, pp. 995-1003: Springer.
- [238] R. S. Sutton, "Learning to predict by the methods of temporal differences," *Machine learning*, vol. 3, no. 1, pp. 9-44, 1988.
- [239] X. Xu, "A sparse kernel-based least-squares temporal difference algorithm for reinforcement learning," in *International Conference on Natural Computation*, 2006, pp. 47-56: Springer.
- [240] B. Deokar and A. Hazarnis, "Intrusion detection system using log files and reinforcement learning," *International Journal of Computer Applications*, vol. 45, no. 19, pp. 28-35, 2012.
- [241] R. Baumann, K. M. Malik, A. Javed, A. Ball, B. Kujawa, and H. Malik, "Voice spoofing detection corpus for single and multi-order audio replays," *Computer Speech Language*, vol. 65, p. 101132, 2021.
- [242] A. R. Gonçalves, R. P. Violato, P. Korshunov, S. Marcel, and F. O. Simoes, "On the generalization of fused systems in voice presentation attack detection," in *2017 International conference of the biometrics special interest group (BIOSIG)*, 2017, pp. 1-5: IEEE.