



A literature review and perspectives in deepfakes: generation, detection, and applications

Deepak Dagar¹ · Dinesh Kumar Vishwakarma¹ 

Received: 2 February 2022 / Revised: 2 June 2022 / Accepted: 13 June 2022 / Published online: 23 July 2022
© The Author(s), under exclusive licence to Springer-Verlag London Ltd., part of Springer Nature 2022

Abstract

In the last few years, with the advancement of deep learning methods, especially Generative Adversarial Networks (GANs) and Variational Auto-encoders (VAEs), fabricated content has become more realistic and believable to the naked eye. Deepfake is one such emerging technology that allows the creation of highly realistic, believable synthetic content. On the one hand, Deepfake has paved the way for highly advanced applications in various fields like advertising, creative arts, and film productions. On the other hand, it poses a threat to various Multimedia Information Retrieval Systems (MIPR) such as face recognition and speech recognition systems and has more significant societal implications in spreading misleading information. This paper aims to assist an individual in understanding the deepfake technology (along with its application), current state-of-the-art methods and gives an idea about the future pathway of this technology. In this paper, we have presented a comprehensive literature survey on the application of deepfakes, followed by discussions on state-of-the-art methods for deepfake generation and detection for three media: Image, Video, and Audio. Next, we have extensively discussed the architectural components and dataset used for various methods of deepfakes. Furthermore, we discuss the various limitations and open challenges of deepfakes to identify the research gaps in this field. Finally, discuss the conclusion and future directions to explore the potential of this technology in the coming years.

Keywords Deepfake · Face-swap · Face reenactment · Lip-sync · Body-puppetry · Speech synthesis · Deep learning

1 Introduction

Nowadays, there has been an uproar of information dissemination in the form of images, video, and audio on the web. An earlier attempt that can be traced in the manipulation field was in 1865, with an iconic photograph of then US president Abraham Lincoln where his face was swapped [1]. Deepfake is the current state-of-the-art in the image, video, and audio manipulation.

Deepfake word is composed of two words, “Deep” and “fake,” which means the fake media [1] that has been created using a deep neural network, a branch of machine learning. Fake media created by this technology appear so realistic (Fig. 1) and believable that it is difficult to identify as counterfeit to the naked eye. This

word became famous when, in 2017, a Reddit user with the name “deepfakes” created pornographic content with a swapped face of a celebrity and posed it online.¹ Since then, it has become one of the hot topics, and there has been a lot of research going on in recent times.

Few deepfake detection challenges are going on to set up the benchmark for deepfake detection. Facebook Inc., in collaboration with Microsoft, AWS (amazon web services), and partnership on AI committee, has created the deepfake detection challenge² (DDC) to encourage the researchers to detect fake and manipulated media. ASV (Automatic Speaker Verification) spoofing Challenge³ is another challenge to detect AI-driven counterfeit voices.

Deepfake has become a buzzword for its versatile application to create different manipulations, especially in images and videos. It has opened the door for various exciting applications in multiple fields such as advertising, film production,

✉ Dinesh Kumar Vishwakarma
dvishwakarma@gmail.com

¹ Biometric Research Laboratory, Department of Information Technology, Delhi Technological University, Delhi 110042, India

¹ AI-Assisted Fake Porn Is Here(vice.com).

² Deepfake Detection Challenge | Kaggle.

³ ASVspoof.

Fig. 1 Example of a Deepfake [2]

creative arts and video games. On the other hand, it poses a threat to various societal problems like manipulating public opinion during elections, undermining journalism, eroding the trust in institutions, committing fraud, creating a false narrative & massive civil unrest and jeopardizing the nation's security. To an individual, it can discredit, blackmail, and harass people to an unimaginative height and may ruin their lives and careers.

The rise of deepfake has also become a potential threat to various multimedia information retrieval systems (MIPR) such as face recognition, speech recognition, biometric and gait analysis systems. Information retrieval systems look for the intended semantic knowledge from multimedia sources (e.g., databases). As the data for various identities are freely available on the internet nowadays, such information allows for creating an intended deepfake mask or audio (using readily available deepfake software), which can persuade the MIPR systems to a fake identity. Moreover, as the manufactured multimedia information freely circulates on the internet, it would dilute the empirical evidence, factual information and critical features extracted by the MIPR systems, which may further create false interpretation of the reality.

1.1 Application of deepfakes

The intention behind using the technology makes it beneficial or threatening to society (Fig. 2).

1.1.1 Beneficial use of technology

This technology offers many benefits if used with the right intention, and it may include the following fields:

Education Deepfake gives a plethora of opportunities for educators with the way they can impart education. For example, videos of historical personalities like Mahatma Gandhi or

Nelson Mandela teaching about themselves and their work. In 2018, a circulated deepfake video of Barack Obama wherein he warns about the dangers of deepfakes was an apt example [3]. Teaching in the above manner makes study compelling for the students.

Entertainment Deepfake has also contributed to entertainment purposes in video dubbing in other languages, memes, GIFs, animating dead or cartoon characters, and special effects in the movies [4]. This industry will indeed leverage advanced features of this technology in the coming times.

Expression Deepfake technology allows people suffering from disabilities, such as ALS (Amyotrophic Lateral Sclerosis, the patient has difficulty speaking and communicating) to express themselves through their deepfake video. Deepfake also allows one to have an avatar experience through virtual engagement that might be impossible to have physically, e.g., in video games [3]. Deepfake can be used for social communication, such as speech, to reach an audience that understands a different dialect than the speaker's natural language [5].

Innovation Deepfake has been used a great deal to attract customers for a brand. For example, Reuters demonstrated the use of an AI-generated presenter-led sports news feed.⁴ In the fashion retail industry, deepfake allows customers to turn into models (virtually) to try out new apparels.⁵ A Japanese company named Data Grid is already using an AI-generated virtual model for advertising [5]. In the future, deepfakes will play a huge role in producing stimulating effects through advertising and branding.

⁴ Synthesia Case studies: Reuters.

⁵ Digital Doubles: The Deepfake Tech Nourishing New Wave Retail (forbes.com).

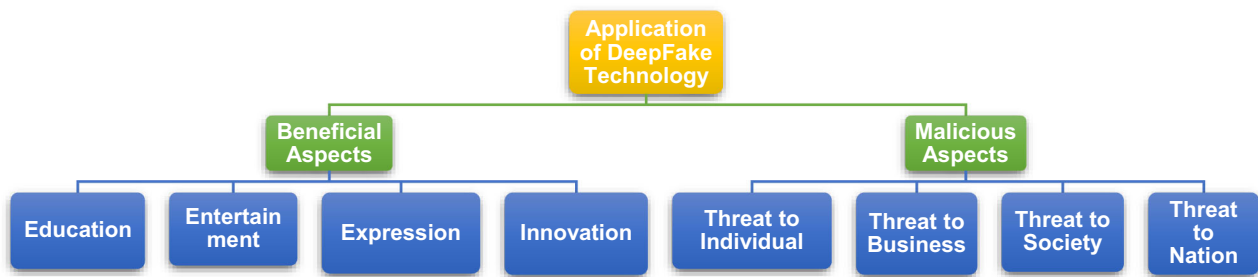


Fig. 2 Application of Deepfake

1.1.2 Malicious use of technology

The real danger of this technology lies in the different ways it can be misused and the large-scale impact it can have, courtesy of being misused. Here are some of the threats that it may create:

Threat to individuals Deepfake holds excellent potential for inflicting tangible harm, psychological stress, and physical pain. A fraudster may use deepfake to extract something of value for inflicting damage. To prevent the release of such deepfake, the victim provides money, personal banking details and business secrets [3]. Also, the most common form of exploitation is in the form of deepfake pornographic videos. One can victimize the individual to any violent or humiliating act to gratify their wants. At the workplace, deepfake videos can be used depicting someone indulging in anti-social actions or behavior like damaging property, hurling abuse or making racist remarks toward a co-worker. Also, deepfake audio or video can be used as evidence for harassment or sexual abuse, which can cause severe repercussions for an individual's career perspective and future aspirations [3].

Threat to the business Deepfake technology allows anyone to impersonate voices of different identities like the Business leader and CEO to incur fraud. A recent example of it, in March 2019, the CEO of a UK-based corporation was asked by his boss impersonated voice to supply \$243,000 to a supplier [6]. Also, a deepfake can manipulate the market quickly if rumors back a fake media, and someone may lose or make a considerable profit [7]. Deepfake ads could be used to defame a company's product and lower its brand value, which could be disastrous for a brand [8]. Deepfake can cause significant harm to an organization's reputation and future perspective in the market and can give undue advantage to its competitor [3].

Threat to society Deepfake can have a substantial societal impact considering its realism and fast propagation through

different social media networks. The rise of deepfake would give tough times to people comprehending what is real and fake, which undermines the very existence of journalism. The Trust of the people in the institutions will erode as this technology prevails. Even if later the video proved fake, it cannot reverse its effect entirely. Deepfake acts as a powerful tool to spread misinformation among the masses on a large scale, creating chaos and panic and division in society. In the worst case, it can lead to a civil war in the community [3].

Threat to nations Deepfake can affect national and international relations; it can sour bilateral ties whose impact may last generations. Deepfake gives the option to external entities to influence the democratic process of a nation [3], which could trigger substantial civil unrest, and protests can turn violent, jeopardizing the nation's security. The well-timed circulation of deepfake can sway elections in a democratic country. The rise of deep fakes could dilute discussions and debates over the policies; empirical evidence would get enmeshed with doctored data. The advancement of such technology dilutes information, putting the speaker's credibility at stake [3].

1.2 Scope and contribution of the surveys

This paper primarily focuses on the application and technical aspects of the deepfake for three media. Technical aspects include deepfake creation and detection methods and datasets proposed from 2019 onwards. Furthermore, we have considered the loosely defined definition of deepfake, where fake media is created by the deep neural networks (DNNs). This paper has comprehensively reviewed existing literature for visual and audio deepfake. The **contribution of this survey** is summarized below:

- Categorize the application of deepfake initially into two categories, which further expand to give a detailed description with different scenarios about the usefulness and misuse of this technology.

- Compare the various state-of-the-art surveys and identify the gaps this review literature aims to address.
- Segregation of detection and generation methods into their relevant categories. Also, detailed descriptions of various generation and detection methods in tabular forms.
- Discuss the various architecture components and datasets used in the generation and the detection methods.
- Outline the various open-source tools used for deepfake creation for multiple platforms and the type of manipulation they do.
- Identify the limitations of the existing generation and detection methods, which need to be addressed by the research community.
- Present the future research trend and pathway of this technology in the coming years.

1.3 Organization of the review

Various research papers have extensively explored the deepfakes, categorizing methods at multiple levels. Tolsana et al. [9] have divided the deepfake techniques into four categories, limited to image and video media without explaining the detection mechanism based on the feature representation. Mirsky et al. [4] have focused on face replacement and face reenactment in great depth. Masood et al. [10] have considered the audio deepfake, but the categorization of detection methods is coarse and not refined further. Earlier literature has assisted in refining the categorization of generation methods for visual and audio media which helps to categorize procedures covering the entire range of techniques. Tolsana et al. [9] help arrive at the categorization for generation methods. Juefei-Xu et al. [11] give insight into feature representation based on detection traces, and also Yu et al. [12] provide an idea for the video media feature representation. These two manuscripts helped arrive at a feature representation that helps refine feature representation based on traces used for detection.

This review paper is segregated into six parts. The first part (Sect. 1) discussed deepfake technology and its application. The subsequent section (Sect. 2) will discuss the deepfake techniques: deepfake generation and detection. Each category contains state-of-the-art methods that have been proposed from 2019 onwards. Subsequently, this paper has presented the popular datasets in the tabular form (Sect. 3). Proposed methods contain various architectural components for their models; this paper briefly introduces such components and tools for deepfake creation (Sect. 4). Next, this paper has highlighted the various limitations of generation and detection methods (Sect. 5). Finally, we end the discussion with the conclusion and future pathway this technology may take in the coming years (Sect. 6).

2 Deepfake techniques

The two major algorithm classes in this area are deepfake generation and detection.

2.1 Deepfake generation

Based on the generation process of different media, the generation mechanism has been divided into visual deepfake, including image and video media, and audio deepfake generation. Figure 3 presents the further categorization of deepfake generation methods for these two categories.

2.1.1 Visual deepfake generation

Visual deepfake generation involves the generation mechanism for image and video media. This section will cover various categories under which manipulation can take place.

Identity swap (IS) The identity/face of the source/frame is transferred to the target image/frame in this type of manipulation. Face-swap is the other name for it. The face-swap [13] and fake app [14] software made it easier for anyone to generate face-swap. Their typical approach for IS is based on a pair of auto-encoder and decoder architecture. For this process, encoder–decoder pairs are required, where the encoder converts the images into their latent representation, while the decoder reconstructs the image back from the latent representation. During the training, encoder–decoder pair is required for each image for the model to learn its embedding, where encoder weights are shared. Once training is complete, the decoder is interchanged during the generation stage, and the decoder of the source image and encoder of shared weights are used to generate the target results (Fig. 4).

The researcher proposed various sophisticated algorithms over the years. The first well-known Identity swap is FaceSwap [15], used to develop the FaceForensics ++ dataset [16]. The traditional method uses 3D morphable models, and facial textures are replaced with the estimated 3D model's geometry with the target image. Dale et al. [17] model has been one of the old approaches that uses the multi-linear model to track the facial performance in both videos and then use 3D geometry to warp the faces. Nowadays, IS architecture uses DNN, which usually uses two modules; one uses latent space for disentanglement of identity from other attributes and then the other module transfers and refines the identity from source to target. Faceshifter [18] manipulation comprises two stages. In the first stage, the method extracts and embeds the target attributes thoroughly and adaptively, and in the second stage, the network recovers the anomalies region in a self-supervised manner. Most of the recent IS methods are subject agnostic, where once the model gets trained, it can be applied to any new faces without re-training

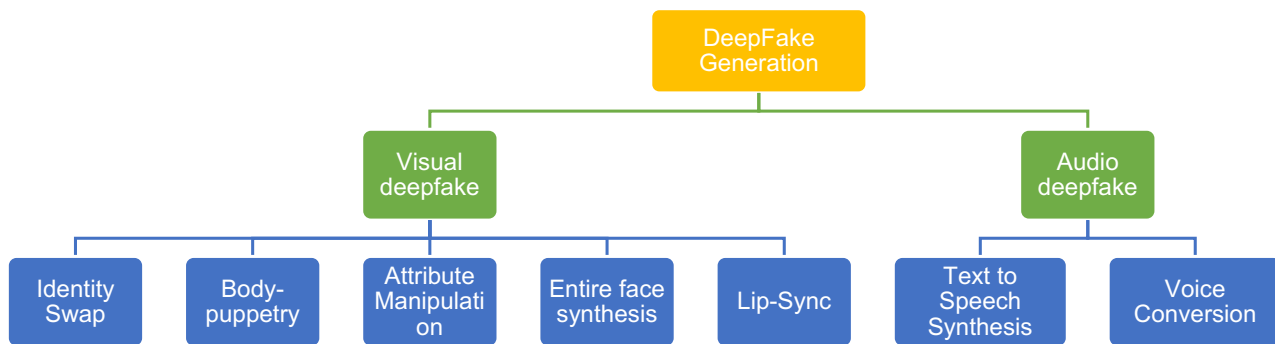
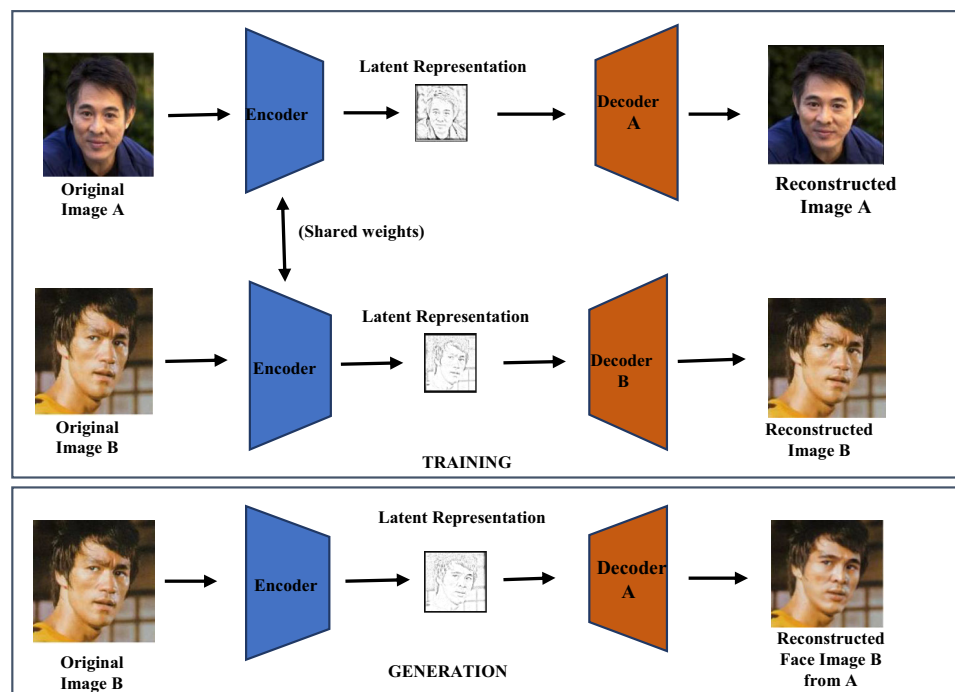


Fig. 3 Classification of deepfake generation techniques

Fig. 4 Identity Swap generation model using auto-encoder & decoder



on subject-specific targets. Table 1 describes recent popular identity swap methods.

Body puppetry (BP, aka reenactment) Body puppetry (aka reenactment) deepfake is where the source derives the content of the target; it can be facial gestures, eye and head movements, or different body poses. Face reenactment is its subset, where the facial attributes are derived. It is significantly used for post-production editing of movies or short videos [25]. Most of the time, the target content is derived either from some source media in the form of images/frames using landmark key points, 3D morphable models, skeletons or any other mapping method. Chan et al. [25] proposed a method to transfer dance moves from the source to the target using intermediate pose Skelton transfer and predicting the two consecutive frames to produce coherent results. However, the samples cannot generate realistic poses, especially

at the body's joints. Thies et al. [26] proposed the famous Face2Face approach that allows for the real-time reenactment of facial expression, where source facial expressions are tracked using a dense photometric consistency measure, and then a transfer function exploits the deformation transfer in semantic space. Many techniques [27, 28, 29] require multiple input images of the source samples, while few methods require few instances [30, 31] or even a single [32] to generate results. Many methods have the limitation that they can synthesize one attribute at a time and apply it to only low-resolution images. FaceSwapNet [33] resolves this issue using two modules: landmark swapper and landmark-guided generator to generate the face expression enacted photo-realistic image. Table 2 lists the popular body puppetry with their analysis in various fields. Some other methods [34, 35, 36, 37, 38] have also been proposed over the years. Most of the reenactment has been done on dance poses and facial

Table 1 Overview of popular identity swap methods

Ref	Year	Model description	Subject agnostic	Dataset used	Worst performance	Best performance	Architectural components	Output resolution	O/P media		Limitation
									Img	Vid	
[19]	2019	Proposed a framework based on a novel recurrent neural network for both face swapping and facial reenactment for a pair of the subject without the training on the subject-specific data	✓	IJB-C dataset, FaceForensics +, LFW, Figaro	SSIM = 0.51, Euler Score = ~ 3.21, Landmark score = ~ 24.5, Verification Score = 0.38	SSIM = 0.54, Euler Score = ~ 2.49, Landmark score = ~ 22.2, Verification Score = 0.38	GANs	128 × 128 & 256 × 256	•	•	The method fails to capture facial expressions & blurriness in image texture fully. Identity and texture quality degrade in case of significant angular differences
[18]	2020	Employed a two-stage GAN framework (called faceshifter) for a face-swapping algorithm. In the first stage, AEI-NET is used to extract & embed target attributes adaptively. In the second stage, (HEAR-Net) attributes are further refined by recovering face anomalies in a self-supervised manner	✓	FaceForensics ++, CelebA-HQ, FFHQ, VGGFace	ID retrieval = 97.38, pose error = 2.96, expression error = 2.06, Human evaluation(identity) = 52.9, Human evaluation(attr.) = 62.1, Human evaluation(realism) = 82.9		Encoder-Decoder	256 × 256	•		Stripped artifacts of the generated samples
[20]	2020	Employed ID Injective Module (IIM) enables identity transfer from source to target image, extending the architecture from specific to arbitrary face manipulation	✓	VGGFace2, FaceForensics + & CelebMask-HQ	Accuracy rate (ID retrieval) = 73.64%, Pose Estimation (L2 Distance) = 1.22	Accuracy rate (ID retrieval) = 96.57%, Pose Estimation (L2 Distance) = 2.47	GAN + Encoder-Decoder	224 × 224	•		Blurriness at the edges

Table 1 (continued)

Ref	Year	Model description	Subject agnostic	Dataset used	Worst performance	Best performance	Architectural components	Output resolution	O/P media		Limitation
									Img	Vid	
[21]	2021	Proposed one-shot learning methods for face swapping, where a hierarchical face encoder encodes face representation in new extended latent space to maintain facial details. A second module transfers the identity from the source to the target by nonlinear trajectory. Finally, the synthesis mechanism of styleGAN2 swaps faces	✓	CelebA, CelebA-HQ, FFHQ, FaceForensics + +	ID similarity = 0.5014, ID retrieval = 90.83, pose = 3.58, expression = 2.87, FID = 10.16	ID similarity = 0.5014, ID retrieval = 90.83, pose = 2.64, expression = 2.96, FID = 10.16	GANs	256 × 256, 1024 × 1024	•	•	Fuzzy detailing at the edges and corners
[22]	2021	Proposed a GAN framework that allows specific attributes control in a high-fidelity face-swapping process. For this, they have used a pose expression (PE) block, which uses a feature-wise boundary map to correct the pose and expression of the manufactured face, along with a discriminator which keeps weak supervision on them. Also, it uses a perceptual loss that leverages the discriminator to help preserve the facial attributes like skin color, illumination and occlusion	✓	FF-HQ, CelebA-HQ, VGGFace2 & FF + +	Identity retrieval(id) = 0.98, DIPD = 0.28, pose dist.(pose) = 2.43, expression dist.(exp) = 2.19, SSIM = 0.77, Avg. endpoint error(AEE) = 1.50, FEW = 0.22, FID = 30.31, SSIM = 0.78, MMS = 0.08	Identity retrieval(id) = 0.98, DIPD = 0.28, pose dist.(pose) = 2.43, expression dist.(exp) = 2.19, SSIM = 0.77, Avg. endpoint error(AEE) = 1.50, FEW = 0.22, FID = 19.72, SSIM = 0.81, MMS = 0.90	GANs	256 × 256	•	•	Visual inconsistencies and blurry artifacts at the boundaries

Table 1 (continued)

Ref	Year	Model description	Subject agnostic	Dataset used	Worst performance	Best performance	Architectural components	Output resolution	O/P media		Limitation
									Img	Vid	
[23]	2021	Unified framework for high fidelity for face-swap and face reenactment using disentanglement of 3D poses shape and different expression factors and then recombining for various tasks	✓	FaceForensics ++, Celeb-DF v2, voxCeleb2	ID Similarity = ~ 0.62, Expression Error = ~ 1.71, Pose Error = ~ 2.62, SSIM = ~ 0.90, Human Deceive rate = ~ 0.35	ID Similarity = ~ 0.80, Expression Error = ~ 1.67, Pose Error = ~ 1.97, SSIM = ~ 0.90, Human Deceive rate = ~ 0.53	CNNs	256 × 256	•		Blending inconsistencies in skin tones, wrinkles, 3D warping error, and contour errors in case of occlusion objects
[24]	2021	Proposed a unified framework for face swapping and face reenactment for video editing. The framework has a 3D dynamic training sample selection mechanism, and the optical flow loss constraint achieves temporal coherence through barycentric coordinate interpolation. In addition, a region-aware conditional normalization layer achieves more realistic results	✓	VoxCeleb2	FID = 34.1, Identity error = 0.26, pose error = (1.12, 1.31, 0.65), expression error = 2.31		GANs	64 × 64	•		Few generated samples are blurry

Table 2 Overview of popular body puppetry methods

Ref.	Year	Model description	Subject Agnostic	Attribute transfer (reenactment)	Dataset used	Worst performance	Best performance	Architectural components	Output resolution	O/P media		Limitations
										Img/frame	Video	
[25]	2019	Employed a motion transfer method that learns the pose mapping and transfers the appearance to the target, predicting two consecutive frames to produce coherent results for the synthesis	✗	Dance poses	YouTube Videos	Human Evaluation = 53.9%, SSIM = 0.816, LPIPS = 0.050	Human Evaluation = 58.8%, SSIM = 0.838, LPIPS = 0.036	GAN	1920 × 1080		•	Model struggles to generate good illustrations for extreme poses. Sensitive to the variation in the scene lighting
[27]	2019	Network combining the traditional graphics pipeline with learnable rendering components through feature maps to produce a photo-realistic image re-synthesis content	✗	Facial Expression	Synthetic objects from artec3d website	Mean Square Error (MSE) = ~ 1.5	Mean Square Error (MSE) = ~ 0.4	Encoder-decoder	512 × 512		•	Reliance on a geometry proxy. Requires to re-train the model for new samples
[39]	2020	Two main components, Dense Mapping Network (DMN) and Editing Behavior Simulated Training (EBST), are employed for diverse, interactive face manipulation	•	Facial Expression & Style	CelebA-HQ, CelebA	FID = 37.55%, A.C.Acc. = 68.1%, H. eval. = 28%, F.V.A = 76.41%, seg. = 92.31%	FID = 37.14%, A.V.Acc. = 89.5%, H.eval = 44%, F.V.A = 76.41%, seg. = 92.34%	GAN	512 × 512	•		Fail to preserve the finer details of the regions

Table 2 (continued)

Ref.	Year	Model description	Subject Agnostic	Attribute transfer (reenactment)	Dataset used	Worst performance	Best performance	Architectural components	Output resolution	O/P media		Limitations
										Img/frame	Video	
[30]	2019	Proposed a few-shot learning-based pre-train model transforming facial landmark positions to generate a realistic-looking head	○	Facial Expression & head pose	VoxCeleb1, VoxCeleb2	FID = 46.1, SSIM = 0.61, CSIM = 0.15, Human Evaluation = 0.62	FID = 29.5, SSIM = 0.74, CSIM = 0.47, Human Evaluation = 0.33	GAN	256 × 256	○	•	Sensitive to the source identity leakage and Lack of gaze adaption
[40]	2019	Employed a progressive pose attention transfer network that smoothly cascaded the input image features and the conditional pose using Pose-attentional Transfer blocks (PATBs)	•	Poses	Deepfashion and Market-1501	SSIM = 0.311, IS = 3.209, M.SSIM = 0.811, mask-IS = 3.773, DS = 0.74, H.eval = 19.14%	SSIM = 0.773, IS = 3.323, M.SSIM = 0.811, mask-IS = 3.773, PCKh = 0.96, DS = 0.976, H.eval = 63.47%	GAN + Encoder-Decoder	128 × 64, 256 × 256	•		Unwanted, distorted and blurry visual artifacts
[31]	2019	Employed a framework vid2vid utilizes the attention mechanism, which uses a few images of the target image to generate videos of unseen using input semantic videos or masks	○	Head poses, dance poses & street scene	CityScapes, ApolloScape, Camvid, FF & YouTube videos	FID = 144.24, Pose Error = 6.01, mIoU = 0.408, Pixel Accuracy = 0.831	FID = 80.44, Pose Error = 6.01, mIoU = 0.408, Pixel Accuracy = 0.831	Encoder + Decoder			•	Does not generalize to unseen domains, and the input is limited to semantic estimations

Table 2 (continued)

Ref.	Year	Model description	Subject Agnostic	Attribute transfer (reenactment)	Dataset used	Worst performance	Best performance	Architectural components	Output resolution	O/P media		Limitations
										Img/frame	Video	
[28]	2019	Employed a neural network that first tracks the 3D motion, transferring it to synthetic renderings to generate corresponding realistic human poses	✗	Dance poses	Videos from Youtube	L2 error 15.67433, SSIM = 0.65328		GAN + Encoder-Decoder	512 × 512		•	Does not generalize for an arbitrary person. Visual artifacts occur mainly at the end-effectors like hands or feet
[41]	2019	Employed a generative network that uses two branches; one branch learns the generation of various poses, and the other improves the temporal coherence of unseen poses	✗	Dance poses	Web videos	P.to.P = 17.2, p.to.S metric = ~ 29(dist), MSE = 16.9, T.C.E = 0.0140	P.to.p = 4.2, p.to.seq = 0(dist), MSE = 15.4, T.C.error = 0.0092	GAN	256 × 256		•	Fail to generate the fine details and also does not change the background
[42]	2019	Leverages two generative models: the first synthesized novel poses, and then the latter fuses the background details, further refining the realism of the video frames	✗	Dance poses	YouTube videos	MSE = 642.9080, SSIM = 08.115, PSNR = 21.3286, Human Evaluation = 29.66%	MSE = 171.3259, SSIM = 0.9352, PSNR = 26.172, H.eval = 41.73%	GAN + Encoder-Decoder	256 × 256, 512 × 512		•	Limited to the generation of static background frames

Table 2 (continued)

Ref.	Year	Model description	Subject Agnostic	Attribute transfer (reenactment)	Dataset used	Worst performance	Best performance	Architectural components	Output resolution	O/P media		Limitations
										Img/frame	Video	
[43]	2020	A two-stage pipeline in which the first extracts the facial expression and attributes (Action Units), and the latter transfers those attributes onto the image	•	Facial expression poses	VoxCeleb, CelebA	Accuracy = 59.88, F-score = 0.3759	Accuracy = 62.86, F-score = 0.4185	GAN	128 × 128	•		Limited resolution and lack of diversification of the output image poses
[44]	2019	Network that first produces the intermediate generation of the correct and detailed poses, conditioned to create temporally coherent photo-realistic video frames	•	Poses	Fashion & Tai-Chit	P _{loss} = 0.5960, FID = 75.44, AKD = 3.77	P _{loss} = 0.2811, FID = 13.09, AKD = 1.36	GAN + Encoder-Decoder	256 × 256		•	Limited to the single source video generation
[33]	2019	FaceSwapNet framework uses two modules, landmark swapper, and landmark-guided generator, to generate the face expression enacted photo-realistic image	•	Facial Expression	RaFD	SSIM = 0.659, FID = 13.26, Human Evaluation = 74.9%	SSIM = 0.711, FID = 11.67, Human Evaluation = 74.9%	GAN + Encoder-Decoder	256 × 256	•		Visual inconsistencies of the generated samples

Table 2 (continued)

Ref.	Year	Model description	Subject Agnostic	Attribute transfer (reenactment)	Dataset used	Worst performance	Best performance	Architectural components	Output resolution	O/P media		Limitations
										Img/frame	Video	
[29]	2021	Framework uses a sequential generator and an ad-hoc discriminator to capture the complex non-rigid facial movement to synthesize a temporally consistent video	✗	Facial expression, head pose & eye gaze	Head2Head(Self-built), FaceForensics++	Av. Exp. Dist.(AED) = 1.509, Av. Rot. Dist.(ARD) = 0.83, DAI = 8.86, FID = 1242, Max. Mean Discrepancy(MMD ²) = 3915, Av. Eye land. Dist.(AELD) = 0.71		GANs	256 × 256		•	Data-intensive model and also computationally expensive
[32]	2021	Composed of three networks, where the first separates the pose and other frame generation parts, the second converts the posture and target id to a masked frame, and the third network does the further refinement	•	Dance poses	Multi-Human Parsing(MHPv2), CIHP, DFDC	SSBS = 0.902, SSIS = 0.218, Dense Pose Binary similarity(DPBS) = 0.928, Dense Pose Index Similarity(DPIS) = 0.5, LPIPS = 0.375, SSIM = 0.116, FID = 83.95		CNNs	512 × 320	•	•	Does not produce coherent results, appearance is not preserved, and edges are pixelated

expressions. There has been a considerable improvement in the quality of dance poses generated samples, but still, they are far from appearing realistic.

Lip-syncing (LS) This category of video manipulation involves synthesizing the mouth region of a target identity consistent with the arbitrary input audio. Lip movement and the corresponding expression are key elements to convey the information effectively. Usually, Influential leader's deep-fakes are developed, as their audios, videos, and images are readily available, and their generated samples create more impact. Suwajanakorn et al. [45] created one of the first well-known lip-sync of ex-President Obama from the audio using a recurrent neural network to map raw audio features to different mouth textures. However, a new identity requires the model to be trained again. Earlier lip-sync methods [46, 45] construct 3D talking face models for a specific by animating 3D face meshes of the particular chosen subject, and such techniques are hard to scale to arbitrary identities. However, as the technology evolves, they disentangle audio-visual representation, allowing methods [47, 48] to use few subject samples to generate results. Real-time reenactment of audio has also become possible nowadays. Jamaluddin et al. [49] proposed a real-time cross-modal self-supervision model for synthesizing talking heads. They employ encoder–decoder multi-stream CNN, which uses a joint embedding of still images and audio to generate lip-synched video frames in real time. However, the model lacks the synthesis of real-time emotional facial expressions. There has been significantly less work for lip-sync manipulation, and also, methods have difficulties generalizing to any arbitrary identity. Table 3 gives an overview of popular lip-sync methods.

Attribute manipulation (AM) Attributes like expressions, hair, eyes, the color of skin, age, gender, mustache, etc. that are manipulated in an image fall into the manipulation category.

Generally, attribute manipulation methods employ either encoder–decoder (ED) or a combination of ED and GANs with a conditioned attribute. ED-based model decodes the latent representation of attributes in a latent representation. A relationship is established between latent representation and attribute independent editing, which allows the independent attribute manipulation without the identity information loss, leading to a distorted or over-smooth generation of the results.

StarGAN and STGAN are classic examples. Earlier domains used to do the image-to-image translation between two domains, which was time-consuming, but StarGAN's [50] approach uses multi-domain image-to-image translation using a single model. It allows the training of multiple datasets of different domains within the same network. However, the model can only produce a limited number of

expressions despite such flexibility. To address this limitation, Albert et al. [51] novel GAN-based model named ganimation uses a weakly supervised attention mechanism that takes annotated facial action units (AU) as input and generates a broader range of expressions. Encoder-decoder and GAN architecture has bottleneck layers, which results in blurry and low-quality results and adding skip connection to overcome these, results in weakened attribute manipulation. Ming et al. [52] proposed STGAN using selective transfer manipulation that incorporates specific target units into the encoder-decoder model, changing target face attributes. While manipulating, Guim et al. [24] use latent space and conditional attribute representation, which helps regenerate the image by modifying the required attribute. Table 4 gives an overview of some popular method of attribute manipulation methods. Some other techniques [53, 54], which have been proposed recently, manipulate the style of an image using StyleGAN [55] latent space.

Entire image synthesis (EIS) The powerful GANs create entire non-existent images with high realism. ProGAN [67] and StyleGAN1 [55] have leveraged the power of GANs to create highly realistic synthetic high-resolution images. Terro et al. [67] proposed the ProGAN methodology, which allows for generating high-resolution images progressively by adding the number of layers gradually with the training. They started with a low-resolution image and then gradually increased the layers of the GANs and the image's resolution. The image generated by the model is of high quality but is far from appearing real at times. Another method, StyleGAN1 [55], interpolates the various features such as pose and human identity by disentangling the high-level attributes from the stochastic variation (like freckles, hair) of the generated image in an unsupervised setting. The method enables intuitive, scale-specific control of the synthesis. However, they found several typical artifacts of StyleGAN. To improve the model, they redesigned the architecture StyleGAN2 [68] with the normalization used in the generator. They also adapted the progressive GAN approach by regularizing the mapping of the generator from the latent code to images. Most methods have a hard time finding the trade-off between fidelity and the variety of generated samples. Table 5 gives an overview of popular image synthesis methods.

Summary of the visual deepfake generation methods Several deepfake generation methods have been proposed over the years. Realism has increased over the past few years, and now the naked have difficulty identifying. Despite being so realistic, in the case of multi-modal data like the video at specific frames, the visual inconsistencies appear, causing flickering and jitteriness. Lip-sync and Body-puppetry, especially in the case of poses, need some improvements to be more realistic at specific frames.

Table 3 Overview of popular lip-sync methods

Ref.	Year	Model description	Subject agnostic	Dataset used	Worst performance	Best performance	Architectural components	Output resolution	Limitations
[56]	2019	Employed a framework called Disentangled Audio-Visual System (DAVS) that generates high-quality talking video frames by integrating the disentangled speech and Identity space representation using associative and adversarial training processes	•	LRW, MS-Celeb-1 M, VoxCeleb	PSNR = 25.4, SSIM = 0.859, Human Evaluation(realistic) = 51.5%, Human Evaluation(Lip-Audio sync) = 72.3%	PSNR = 26.8, SSIM = 0.884, Human Evaluation(realistic) = 87.8%, Human Evaluation(Lip-Audio sync) = 88.4%	GAN + Encoder-Decoder	256 × 256	Method lacks the synthesis of emotional facial expressions
[46]	2019	Employed a model based on a dynamic programming strategy for talking head video that allows you to edit, insert and delete the text in an existing transcript	✗	DF-TIMIT videos also taken from YouTube	RMSE = 0.021, Human Evaluation(real) = 57.1%	RMSE = 0.018, Human Evaluation(real) = 62.1%	RNN + GAN + Encoder-Decoder	512 × 512	The model requires huge time (at least 1 h of video) to produce the best quality results. In addition, occlusion of the lower face region would lead to synthesis artifacts

Table 3 (continued)

Ref.	Year	Model description	Subject agnostic	Dataset used	Worst performance	Best performance	Architectural components	Output resolution	Limitations
[57]	2019	Employed a method for generative talking videos by converting audio signals to facial landmarks and mapping them to facial appearances in video frames. Also employed an adjustable pixel-wise loss with an attention mechanism and regression-based discriminator for smooth facial movement and audio-visual consistency	•	LRW, GRID, VoxCeleb, TCD and real-world samples from YouTube	LMD(Landmarks Distance) = 1.82, PSNR = 28.78, SSIM = 0.72	LMD(Landmarks Distance) = 1.29, PSNR = 0.83, SSIM = 32.15	RNN + Encoder-Decoder	128 × 128	Limited to the conscious head movement of the identity
[58]	2019	Employed an end-to-end temporal generative model that uses three discriminators. It helps attain a detailed level frame, audio-visual sync, and realistic facial expression synthesis to generate talking heads with facial movements using an audio clip and a still image	•	GRID, TCDM, TIMIT, CREMA-D, LRW	PSNR = 20.107 SSIM = 0.658, CPBD = 0.189, ACD = 2.61 × 10 ⁻⁴ , WER = 58.2%, AV offset = 8, AV Confidence = 1.4	PSNR = 27.1 SSIM = 0.818, CPBD = 0.308, ACD = 1.02 × 10 ⁻⁴ , WER = 23.1%, AV offset = 1, AV Confidence = 7.4	GAN + Encoder-Decoder	96 × 128	Limited output resolution and the methods work for only frontal poses

Table 3 (continued)

Ref.	Year	Model description	Subject agnostic	Dataset used	Worst performance	Best performance	Architectural components	Output resolution	Limitations
[59]	2020	Employed an audio-driven facial reenactment model that inherently maps the audio stream to represent the person talking style using 3D latent space. Then, a novel rendering network uses the representation of latent mapping to produce photo-realistic talking video frames	✗	News-speaker videos are taken from the internet	Human Evaluation(visual) = ~ 60%, Human Evaluation(Audio-visual sync) = ~ 60%		RNN	512 × 512	The method fails when the audio stream contains multiple voices and also when the video contains occluded objects
[60]	2019	Employed subject-independent temporal GAN method that directly maps the audio features into visual features to produce the accurate lip-sync photo-realistic talking heads with facial movements	•	GRID, TCD TIMIT dataset	PSNR = 23.01, SSIM = 0.654, CPBD = 0.252, FDBM = 0.097, ACD = 2.29 × 10 ⁻⁴ , WER = 37.2%, Human Evaluation = 20.22%	PSNR = 27.98, SSIM = 0.844, CPBD = 0.280, FDBM = 0.114, ACD = 1.02 × 10 ⁻⁴ , WER = 25.4%, Human Evaluation = 79.77%	GRU + GAN + Encoder-Decoder	96 × 128	Facial expression is solely derived from the spoken word; it does not represent the speaker's mood
[49]	2019	Proposed a cross-modal self-supervision model of synthesizing talking head that employs encoder-decoder multi-stream CNN, which uses a joint embedding of still images and audio to generate lip-synched video frames in real time	•	LRS2, VoxCeleb2, VGG face	MSE = 705, Embedding distance = 0.434, Retrieval accuracy = 79.7%	MSE = 327, Embedding distance = 0.115, Retrieval accuracy = 83.9%	Encoder-Decoder	109 × 109	Model struggles in making the chins move with the mouth, and it also lacks the synthesis of emotional facial expression

Table 3 (continued)

Ref.	Year	Model description	Subject agnostic	Dataset used	Worst performance	Best performance	Architectural components	Output resolution	Limitations
[48]	2021	Proposed a model to synthesize low-guided high-quality talking head generation consisting of two cascaded modules. The first module contains a generator to produce animate facial attributes like the movement of the mouth, head, eyebrow etc. The second module transforms these animations into more expression details, and the flow-guided mechanism synthesizes the video	•	Videos collected from the YouTube	PSNR = 24.4174, SSIM = 0.84, CPBD = 0.153, Human Visual Evaluation, MSE = 0.0875, LMD = 0.1899, CCA = 0.786		Encoder-decoder	720p ~ 1080p	Method cannot produce temporally coherent results, unnatural speaking style, and the head pose is not extreme enough

Table 3 (continued)

Ref.	Year	Model description	Subject agnostic	Dataset used	Worst performance	Best performance	Architectural components	Output resolution	Limitations
[47]	2021	Proposed a framework for animating talking heads faces from arbitrary audio. First, they convert the video footage into normalized that decouples pose, texture, lighting etc. Secondly, they leverage facial symmetry to approximate skin constancy, isolate spatiotemporal lighting, and then use an auto-regressive approach to stabilize temporal dynamics of the resultant frames	○	GRID, TCD-TIMIT, CREMA-D	SSIM = 0.91, LMD = 1.57, CPBD = 0.25, WER = 18%, Human Eval(Lip-sync) = ~ 2.46, Human Eval(Vis. Quality) = ~ 4.10	SSIM = 0.94, LMD = 0.80, CPBD = 1.57, WER = 18%, Human Eval(Lip-sync) = ~ 2.46, Human Eval(Vis. Quality) = ~ 4.10	Encoder-Decoder	256 × 256	Synthesized heads movement seems out of place at various frames, and also, the method is limited to short videos, and it is a bit slower. Hence cannot be deployed in real-world scenarios

Table 4 Overview of popular Attribute Manipulation methods

Ref	Year	Model description	Subject agnos- tic	Attributes manipulated	Dataset used	Worst performance	Best performance	Architectural components	Output resolution	O/P media		Limitations
										Img/frame	Video	
[61]	2019	Employed attribute classification constraints for manipulating desired attributes, reconstruction loss to preserve the non-manipulated details of an image, and adversarial loss for realistic visual learning of the attribute	✗	Bald, bangs, Black hair, Blond hair, Brown hair, Bushy eyebrows, Eyeglasses, Gender, Mouth open, Moustache, No beard, Pale skin, Age	CelebA & LFW	Facial Attribute Editing Accuracy = ~ 0.35 Attribute Preservation error = ~ 0.10	Facial Attribute Editing Accuracy = ~ 0.95 Attribute Preservation error = ~ 0.01	GAN	64 × 64, 128 × 128, 384 × 384	•		Manipulation is limited to facial attributes only and cannot apply to other general attributes of a person
[62]	2020	InterfaceGAN allows semantic editing of facial attributes by interpreting the latent Space representation learned by the GAN model	✗	Gender, age, expression, eyeglasses, pose	CelebA-HQ	Classification Accu = 75.3%, Correlation matrix = -0.08	Classification Accu = 90.3%, Correlation matrix = 1.0	GAN & SVM	-	•		Method is more inclined to the manipulation of synthetic images

Table 4 (continued)

Ref	Year	Model description	Subject agnos- tic	Attributes manipulated	Dataset used	Worst performance	Best performance	Architectural components	Output resolution	O/P media		Limitations
										Img/frame	Video	
[52]	2019	Employed STGAN having an encoder-decoder network that incorporates the specific transfer units as input for target attributes manipulation while improving the resolution of an image	✗	Bald, Bangs, Black hair, Blonde hair, Busy eyebrows, Eyeglasses, Male, Mouth slightly open, Mustache, No Beard, Pale skin and Young	CelebA	Human Evaluation = 47.6, Attribute generation accuracy = ~ 0.25	Human Evaluation = 69.96, Attribute generation accuracy = ~ 0.85, PSNR/SSIM = 31.67/0.9948	GAN + Encoder-Decoder	384 × 384	•		Model performs poorly for multiple attribute manipulation
[63]	2019	Employed a fully end-to-end convolutional network SC-FEGAN using additional loss that uses free-form user input in the form of masks, color, and sketches as a guide to generate the photo-realistic images	✗	Any desired attribute	CelebA-HQ	PSNR = 27.9131, SSIM = 0.9543, L2 loss = 0.2, LPIPS = 0.0795	PSNR = 31.1687, SSIM = 0.9671, L2 loss = 0.0937, LPIPS = 0.0552	GAN	512 × 512	•		Finer details of the skin texture are not preserved

Table 4 (continued)

Ref	Year	Model description	Subject agnos- tic	Attributes manipulated	Dataset used	Worst performance	Best performance	Architectural components	Output resolution	O/P media Img/frame Video	Limitations
[64]	2020	Employed a framework InterFaceGAN that allows interpreting disentangled face representation in linear subspaces that paves the way for manipulating gender, age, expression, and presence of eyeglasses via subspace projection	✗	Gender, age, expression, Eyeglasses, pose	CelebA-HQ(PGGAN) and FFHQ(StyleGAN)	Re-scoring analysis = -0.24, Identity discrepancy = 0.01, Disentangle-ment analysis = -0.42 Class. Acc. = ~ 0.65	Re-scoring analysis = 0.59, Identity discrepancy = 0.61, Disentangle-ment analysis = 1.00 Class. Acc. = ~ 0.10	GAN	224 × 224	•	-
[51]	2019	Employed a novel GAN based on a weakly supervised attention mechanism that takes annotated facial action units (AU) as input and generates a broader range of expressions	•	Facial Expressions	EmotionNet, RaFD, CelebA	Average Content Distance (ACD) = 0.4, User Preference = 0, Expression Dis-tance(ED) = 4.8	IS = 1.48, ACD = 0.0, User Preference = 87, Expression Dis-tance(ED) = 0.4	GAN	-	•	Lack of adaption for various poses and gaze. The application of the model to the video sequence is still questionable

Table 4 (continued)

Ref	Year	Model description	Subject agnos- tic	Attributes manipulated	Dataset used	Worst performance	Best performance	Architectural components	Output resolution	O/P media		Limitations
										Img/frame	Video	
[65]	2021	Employed a two-stage framework for joint manipulation of pose and facial expression for high-resolution images. In the first stage, the boundary is predicted, and in the second stage, structure and texture are disentangled to generate the desired manipulation.	•	Poses and expression	CelebA-HQ, RaFD, MultiPIE, MVF-HQ	FID = 38.65, Rank-1 Accuracies = 48.2, Rank-1 recognition rates = 60.4%	FID = 12.94, Rank-1 Accuracies = 60.4, Rank-1 recognition rates = 100%	Encoder-Decoder	128 × 128, 512 × 512, 1024 × 1024	•		Method is sensitive to the variation in poses and illumination
		A high-resolution database named MVF-HQ is proposed that contains 120,283 images at resolution 6000 × 4000 for 479 identities										

Table 4 (continued)

Ref	Year	Model description	Subject agnos- tic	Attributes manipulated	Dataset used	Worst performance	Best performance	Architectural components	Output resolution	O/P media		Limitations
										Img/frame	Video	
[66]	2021	Proposed a domain-guided Noise-optimization based on an inversion approach for facial image attribute manipulation. First, they map the image into the latent space, aligning the inverse code with semantic knowledge. Then, a noise optimization mechanism for noise addition to capture high-frequency details like edges and corners, finally, a mask fuses image and local style	✗	Any desired attribute	FFHQ, CelebA-HQ	MSE = 0.0054, RMSE = 0.0731. PSNR = 23.318, UQI = 0.9266, SSIM = 0.9980, MS-SSIM = 0.9998, VIF = 0.9012, Att. Gen Acc. = ~0.8	MSE = 0.0043, RMSE = 0.0570. PSNR = 25.669, UQI = 0.9481, SSIM = 0.9978, MS-SSIM = 0.9998, VIF = 0.9535, Att. Gen Acc. = ~1.0	GANs + Encoder	1024 × 1024	•	Method takes a few minutes to optimize the semantics of the code, while the embedding algorithm is done in seconds. Also, the process is not so attractive for interactive editing of images	

Table 5 Overview of popular Image Synthesis methods

Ref	Year	Model description	Dataset used	Worst performance	Best performance	Architectural components	Output resolution	Limitation
[55]	2019	Implemented an enhanced unsupervised style transfer to disentangle high-level attributes from the stochastic variation. To further quantify the properties of disentanglement and interpolation, they have employed two automated methods, i.e., perceptual path length and linear separability	Real images collected from Flickr web-site (FFHQ dataset)	FID = 5.06, Perceptual path(full) = 283.5, Perceptual path(end) = 285.5 separability scores = 9.88	FID = 4.40, Perceptual path(full) = 217.8, Perceptual path(end) = 195.9 separability scores = 3.79	GAN	1024 × 1024	Limited resolution of the generated samples
[69]	2019	Proposed a self-attention mechanism into GANs framework that models long-range and multi-level dependencies and more refined details from several locations of an image. In addition, spectral normalization stabilizes GANs training and a two-timescale learning rate(TTUR) to speed up the discriminator training	ImageNet2012	FID = 22.98, IS = 43.15	FID = 18.28, IS = 52.32	GAN	128 × 128	Undesirable artifacts are visible at various patches of the image
[70]	2019	Train the GANs on the large-scale ImageNet dataset to study the instabilities and then apply orthogonal regularization to the generator to make it amenable for truncation trick, allowing explicit finer control for the trade-off between variety and fidelity of the image synthesis	ImageNet, JFT-300 M	FID = ~ 39.7, IS = ~ 124.5	FID = ~ 5.7, IS = ~ 298	GAN	128 × 128, 256, 256, 512 × 512	Conditional class synthesis of image
[68]	2020	Redesigned the architecture of StyleGAN1 [55] with the normalization used in the generator, adapting the progressive GAN approach to increase the number of layers as the training progresses and regularizing the mapping of the generator from the latent code to images	FFHQ, LSUN CAR	FID = 2.84, Perceptual Path length (PPL) = 415.5, Precision = 0.678, Recall = 0.492	FID = 2.32, Perceptual Path length (PPL) = 145.0, Precision = 0.689, Recall = 0.514	GAN	256 × 256, 512 × 384, 512 × 512, 512, 1024 × 1024	–

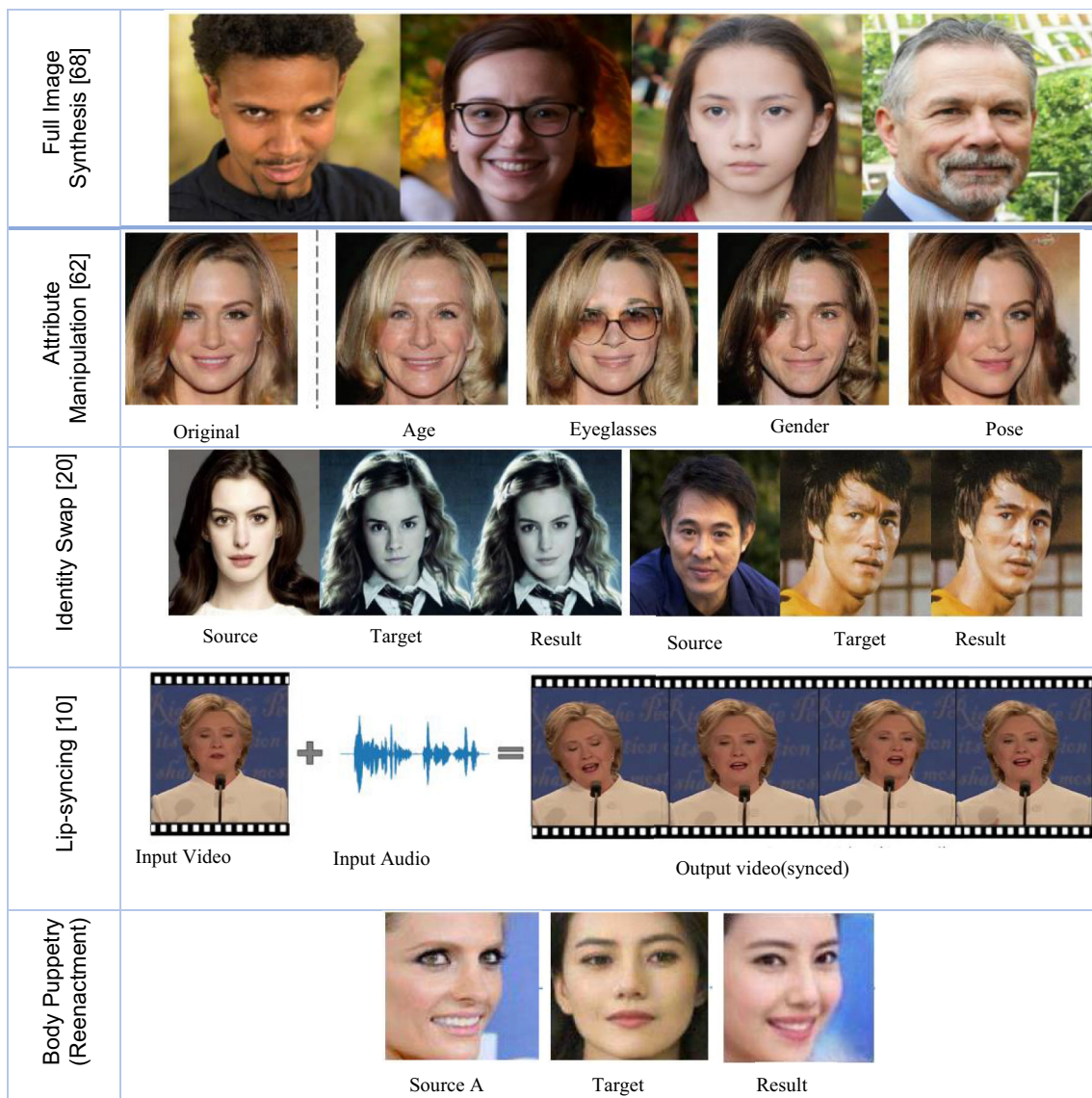


Fig. 5 Types of visual deepfake manipulation

There has been a common trend that we can see that VAEs and GANs have been state-of-the-art generative models used more frequently. VAEs are often utilized for attribute manipulation, where the latent space is exploited to disentangle features and modify the attribute without affecting the other. RNNs like LSTM have been used for sequence data. Evaluating the fidelity of the generated results has always been an issue; most of the authors have adopted objective and subjective evaluation a great deal. For objective evaluation, FID, IS, SSIM, and PSNR have been used more frequently for evaluating the results. For subjective assessment, a human visual score has been adopted where the user is asked to rate the realistic nature of the visual results. Also, methods have a hard time finding a balanced trade-off between fidelity and variety of the generated samples. The CelebA dataset has been used

most frequently, and some others even choose to take the dataset directly from the internet. Among the output resolution of the generated results, we can see that, on average, 256×256 is the most generated output of the models, suggesting that models struggle with high-resolution visual samples. Although few methods like PGGAN [67], and StyleGAN1 [55], can have high resolution and realistic content. The most common limitation that we encounter is that the generated results are not generalizable to other identities; they need to re-train their model for a new identity. Also, the generation mechanism has to go a long way before they provide real-time manipulation with realistic results. There have been few methods for Lip-sync and entire image synthesis among the five categories of manipulation (Fig. 5).

2.1.2 Audio deepfake generation

The deepfake generation mechanism has also extended to audio manipulation. This manipulation form clones someone's voice or generates a speech from the text using audio samples. It just needs someone's audio samples to create the audio of the people, which they have never said. Speech synthesis is divided into two categories:

Text to speech (TTS) TTS synthesis refers to generating speech from the text in someone's voice. Such systems require a database of audio samples of the speaker to train the system; once the system is trained, it can generate the audio sample in the speaker's voice from the input text. TTS systems have become indispensable components of our daily lives. It is extensively used in human-technology interfaces like virtual voice assistants (e.g., Google Assistant, Apple's Siri, Microsoft's Cortana, & Amazon's Alexa), GPS navigation systems, speech-to-speech translation across multiple languages, and screen readers [71]. Such systems are beneficial for visually impaired people to access the technology without any aid. On the other side, the easy availability of TTS synthesis methods can be misused to generate someone's voice without their consent.

Basic pipeline of TTS systems: The input to the model is the text that goes through stages and finally generates an audio waveform (Figure 6) [72]. Different stages have different functions:

Preprocessor: This stage takes text as an input and breaks it into phonemes (linguistic features) based on pronunciation. From the phonemes, phoneme duration, pitch and energy (indicates the magnitude of Mel-spectrograms) are used to train the corresponding components of the systems to get the more natural output of voice [72].

Encoder: An encoder takes the linguistic features and converts them into n -dimensional latent feature embeddings. The speaker encoder also converts the speaker's voice into its embedding, which is concatenated with phonemes to get latent features representation of text and audio [72].

Decoder: At this pipeline stage, the decoder converts the latent feature into an acoustic feature, a Mel-spectrogram representation of audio. Mel-spectrogram is obtained from a short-time Fourier transform by applying a nonlinear transformation to the frequency axis [72].

Vocoder: It converts the Mel-spectrogram representation of audio into waveform. There are different ways to do it. It can be done using Griffin Lim methods or a neural network to map Mel-spectrogram to the waveform. Nowadays, neural network techniques outperform the Griffin Lim method [72].

TTS systems have been developed using two major approaches: concatenative and parametric approaches. In concatenative TTS, the process is based on collecting and fragmenting the good-quality speech recordings into the

units. Then concatenating fragmented units into their corresponding text to generate the speech [74]. However, this approach produces the glitch text for the unseen part of the speech, and a new modification requires an entirely new dataset collection. In parametric TTS, the aim is to use the physical parametric model, usually, a function, to reenact the vocal tract of a human being and then adapt those parameters using recorded speech [74]. Such models allow modification of the voice of speech by tuning the parameters quickly. After the recent advances in deep learning techniques, TTS systems can easily modify the speech and produce a more natural speech. WaveNet [75] was one such kind developed by the researchers at DeepMind in 2016 using raw audio waveforms. The auto-regressive probabilistic model uses the previous audio sample to predict future audio, trained on thousands of audio samples of different subject identities to generate different voices. It produces more natural speech than the various state-of-the-art methods, and the model shows promising results when applied to other audio domains like music. However, WaveNet has a limitation of its prohibitive sequential speed that enables it to generate one audio sample at a time. Hence, it cannot be deployed in real-time settings requiring parallel data generation. Parallel WaveNet [76] was proposed to counter the prohibitive sequential speed of the Base WaveNet model. The model uses the probability distillation method to train the parallel feed-forward network using the sequential trained WaveNet model, generating high-fidelity audio samples at $20 \times$ more speed than the usual WaveNet model.

Deepvoice1 [77] model, based on a deep neural network, is known for its production quality TTS systems. The framework with five building blocks requires few parameters for manual annotation, trains faster, and can work as standalone systems without the pre-existing model. DeepVoice2 [78] was another deep learning model developed in succession for multi-speaker voice systems to counter the limitation of single-speaker voice. The model uses the combination of Deep voice 1 & Tacotron model, where each speaker's embedding is represented in low-dimensional space, and the model's parameters are shared among each other. Finally, deep voice3 [79] is a full sequence-to-sequence conventional acoustic model agnostic of the waveform synthesis methods that have been scaled to over two thousand recordings.

In 2017, Google introduced Tacotron [80], an end-to-end speech synthesis method from a given text. The model can be trained with random initialization using a $\langle \text{text}, \text{audio} \rangle$ pair and output spectrogram. However, Tacotron suffers from the flaw introduced due to the difference between the training and test processes, which affects the synthesis quality and length of the generated frames. Tacotron2 [84] uses an unsupervised GAN model to tackle the problem. The model introduced a training strategy called "random down" and

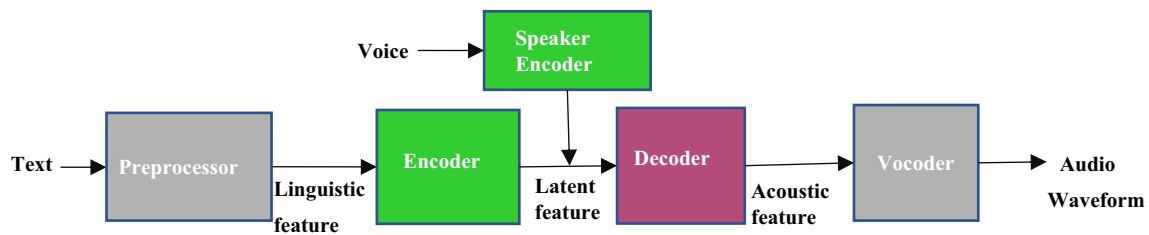


Fig. 6 High-Level diagram of different components used in TTS systems [73]

added the window weights to take care of the minimum values of the attention weight. Table 6 gives an overview of current text-to-speech (TTS) methods.

Voice conversion (VC) Voice conversion is another category of speech synthesis, which deals with converting the source speaker's voice to another form as if uttered in the target speaker's voice while maintaining their linguistic content. This system can be applied as a component in personalized TTS systems for voice conversion [81]. Apart from that, it can be used for impersonating or hiding a person's speech, communication aid for speech-impaired people [82], speech enhancement/accent conversion in language learning and voice dubbing in movies [83]. Voice conversion has been one of the most active research areas in the past few years. Zhang et al. [84] perform non-parallel voice conversion methods by disentangling the linguistic information of the speech from the speaker's characteristics. Voice conversion methods generally require a large dataset for the source and target speaker to learn the mapping. However, few methods have been developed in the last few years that require less data and can even perform voice conversion for unseen voices. Wang et al. [85] propose a StarGAN framework that uses an incremental approach for voice conversion for the unseen dataset. Liu et al. [86] use a transformer network, which performs end-to-end voice conversion, based on three ideas: transformer, context preservation and model adaptation.

Basic voice conversion pipeline: The basic pipeline for VC, shown in Fig. 7, is also called as analysis-mapping-reconstruction pipeline [83]. Linguistic factors, supra-segmental speech factors (prosodic characteristics of speech), and segmental factors characterize the speaker's speech. During the training, mapping is done between the source and target speech features represented in the function form. At the inference time, source speech is given, decomposed into supra-segmental and segmental features by the speech analyzer module, and then features are extracted [83]. Then, in the mapping module, mapping is done between the source speech features and the corresponding target features

learned during the training time. Finally, the reconstruction module synthesizes speech signals [83]. Table 7 briefly describes the Voice Conversion(VC) approaches.

Summary of the audio deepfake generation methods Various methods have been proposed for text-to-speech (TTS) and voice conversion. There are a few standard things about such approaches. CMU artic, LJspeech and VCTK database are the most preferred dataset among the research community for their models. Encoder–decoder architecture components are widely used, given the bottleneck it provides to alter the embedding of the speech to get the desired results. Like visual deepfake, audio deepfake has an issue with evaluating the generated samples; there has not been a consensus among the community about the metric to be used for the evaluation. That is why authors have preferred both subjective and objective assessments to prove their results. MCD, RMSE, and cosine similarity are the most frequently used metrics for objective results. For subjective evaluation, MOS is used, where a user rates the audio on various parameters like naturalness, similarity, emotional, etc. The most significant limitation that the methods suffer is that they fail to produce the natural and emotions contained voice. Also, the technique is not generalizable to various languages, and models also are data-intensive and need much data before generating results. As a result, methods struggle to generate results for low-quality and noisy data samples.

2.2 Deepfake detection

There is an arms race between manipulators and the detector; the detector uses some clue to find the detection, and the manipulator tries to diffuse it next time to make it detection-proof. Generalizing the deepfake technique, robustness against various post-processing operations and interpretability of the detection results are three critical factors for a detector to be deployed in the wild [11].

Visual deepfake and audio deepfake are the two different media for which different detection algorithms have been designed. Based on the clues/traces of feature representation, these two categories have been divided further, as shown in Fig. 8.

Table 6 Overview of different Text-to-Speech methods proposed from 2019 onwards

Methods	Year	Model description	Dataset	Architectural Components	Worst Performance (Evaluation Metrics)	Best Performance (Evaluation Metrics)	Limitations
[87]	2019	Extended the Tacotron system to Japanese speech synthesis systems with a self-attention mechanism to capture long-term information on pitch accent	Japanese Speech Corpus	LSTM + Encoder-Decoder	MOS(naturalness) = 2	MOS(naturalness) = 5	The absence of word-level info in linguistic features and also model lacks real-time speech synthesis
[88]	2019	Proposed an adaptive model for TTS based on the meta-learning concept that uses WaveNet as a core model. The model first learns the speaker's embedding, then fine-tunes the architecture, and then predicts the embedding of the speaker	LibriSpeech & VCTK	Encoder + Deep neural networks	MOS = ~ 4.01, EER = 7.34, Cosine Similarity = 0.10	MOS(naturalness) = ~4.12, EER = 1.85, Cosine Similarity = 0.75	Performance degradation for a low-quality noisy audio sample. The model requires a lot of high-quality training data
[89]	2019	Introduced the unsupervised GAN model that optimized the prosody of the generated speech and incorporated a training strategy called 'random down' to alleviate the cumulative error problem of Tacotron2	Text taken from the web	GANs	Graphical evaluation		Method is limited to the input text length of 1000
[90]	2021	Extending the architecture model of the tacotron by including the normalizing flow into the auto-regressive model. The sequential decoder model takes the input, produces conditioning features for a normalizing flow, and generates output waveforms containing various interdependent samples	Lj speech & a proprietary dataset	Vocoder with Attention mechanism	MCD = 6.87, MSD = 9.24, CER = 9.2, MOS = ~ 3.56	MCD = 4.64, MSD = 11.44, CER = 9.4, MOS = ~ 4.23	Model is computationally complex

Table 6 (continued)

Methods	Year	Model description	Dataset	Architectural Components	Worst Performance (Evaluation Metrics)	Best Performance (Evaluation Metrics)	Limitations
[91]	2021	Proposed a multi-rate attention architecture to cope with the speech generation mechanism's low latency and slower synthesis speed. The encoder uses the parallelizable approach computing the compact multi-headed representation in a streamlined manner and then performing dynamic pooling to restrict the length to produce high-quality content	TTS dataset recorded in a voice production studios	CNNs with Attention Mechanism + •LSTM + Vocoder	MOS = ~ 4.08	MOS = ~ 4.31	-
[92]	2021	Proposed a novel training strategy to capture the association between input text and its corresponding prosody styles using Tacotron-based systems, improving speech's expressiveness. The supervised system training includes a fully differential perceptual and frame reconstruction loss that does not require any reference speech or manual style selection during execution	IEMOCAP, LJ-Speech database	LSTM with Attention Mechanism + Decoder + CNNs	MCD = 7.01, RMSE = 1.53, FD = 15.59, MOS = 3.63	MCD = 6.37, RMSE = 0.94, FD = 13.96, MOS = 4.19	Approach experimentation is limited to the Tacotron systems
[93]	2021	Proposed a framework for a sequence-to-sequence model for concatenative speech synthesis. The encoder has the same design as Tacotron2, while the decoder has been redesigned. The decoder contains an auto-regressive mechanism at both frame-level and phoneme levels that learn the mapping between phone sequences and acoustic using the transition probabilities	Chinese corpus	Encoder + Decoder-LSTM with Attention Mechanism	MSD = 3.70, MCD = 3.64, RMSE = 39.84, CORR = 0.89, UV = 7.31, Preference score = 51.0	MSD = 3.64, MCD = 3.46, RMSE = 37.76, CORR = 0.89, UV = 7, DRMSE, Preference score = 55.0	Method's experimentation is limited to the Chinese language corpus

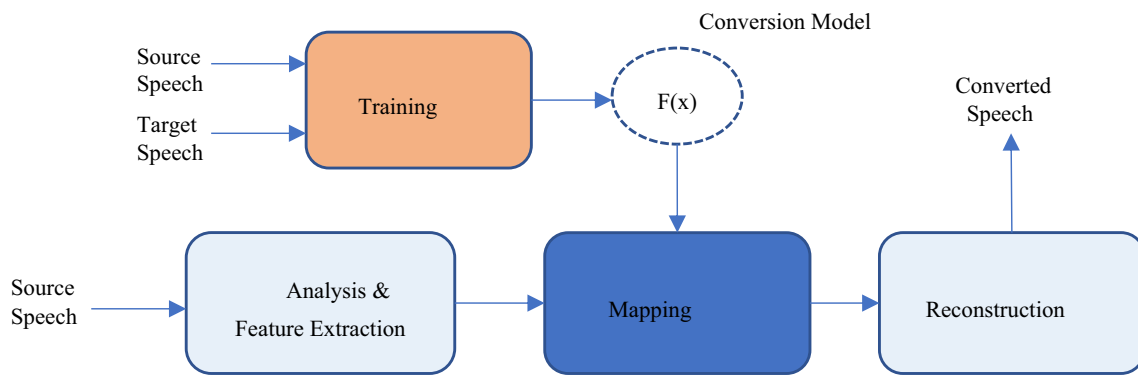


Fig. 7 Typical flow diagram of Voice Conversion system [83]

2.2.1 Deepfake visual detection

Manipulated or manufactured samples have a certain peculiarity in spatial, temporal or frequency domain representation, which the different detection algorithm exploits. The subsequent sections will discuss the various detection techniques along with other domains.

Spatial domain-based detection The change in pixel distribution of manipulated samples can be reflected in the spatial domain properties. This section will discuss various types of spatial domain-based detection methods. Table 8 highlights the recent spatial domain-based detection approaches.

Forensics based detection Generation methods leave certain clues or traces that change the distribution of the samples, which is exploited by the detection methods by analyzing latent features and patterns. Li et al. [105] analyze the subtle distribution of the image statistics in the chrominance components of YCbCr and HSV color spaces, especially in the residual domain for the unseen DNG images detection. Chen et al. [106] use a multi-domain architecture where the features from the RGB domain and the noise vectors (which get added by the external mechanism) are fused to obtain richer robust features.

There are various variants of Forensics features are mentioned below:

- **GAN-Artifacts-based artifacts:** The imperfect design of the GANs leaves some traceable clues, which various researchers use for fake detection investigations. McCloskey et al. [107] utilized the prior knowledge about how the color is treated in GAN and camera models and used this knowledge as a cue to design the network. Methods perform exceptionally for the existing GAN model; however, the model's performance is unclear for new advanced GANs. Yu et al. [108] identify a unique GAN stable fingerprint that persists across different frequencies

and patches of the generated image extracted by a neural network. However, the method fails for post-processing operations like compression, blur etc.

- **PRNU Noise-based detection:** Photo response non-uniformity (PRNU) is a noise-like pattern in the digital image caused by the camera's light sensor. Koopman et al. [109] proposed a method where Co-relation scores are computed between every eight groups of PRNU pattern frames, which serve for deepfake video detection. However, their evaluation is limited to the small dataset. PRNU based are generally low-cost based methods and have a high generalization capability.

Visual-artifact based detection The synthesized or manipulated faces would reveal inconsistencies in the appearance, especially the blending boundaries, landmark points or the shape of the manipulated facial attributes. Even sometimes, the content of the face also seems strange to the rest of the background. Li et al. [110] use the face warping artifact as a clue caused due to blending operation to match the configuration of the source face. Unfortunately, this affine warping operation leaves the artifact due to resolution inconsistency between the face and the surrounding area. The method is more robust than other existing methods, but there is still room for improvement.

Visual-Artifact based methods can achieve better generalization as they pay more attention directed to the specific artifacts. Li et al. [111] propose a generalized detector that uses face X-ray and blending boundaries for detection purposes. However, such a method fails for entirely synthetic images, indicating the blending operation's absence. Also, adversarial samples can be designed to bypass the detection mechanism by curbing such manipulation clues. Matern et al. [112] also exploited the visual artifacts like the difference in eye color, inconsistent illumination, missing teeth areas, etc.; once the algorithm extracts the artifacts, they are used for classification. These methods can localize the manipulation easily, as their detection is based on specific localized

Table 7 Overview of various Voice Conversion(VC) methods proposed from 2019 onwards

Ref	Year	Model description	Dataset	Architectural components	Worst Performance (Evaluation Metrics)	Best Performance (Evaluation Metrics)	Limitations
[94]	2019	Proposed a sequence-to-sequence learning-based method (ATTS2S-VC) that uses context preservation and attention mechanism for voice conversion tasks	CMU Arctic Database	Encoder-Decoder	Preference test(naturalness) = ~ 30% Preference test(speaker similarity) = ~ 25%	Preference test(naturalness) = ~ 60% Preference test(speaker similarity) = ~ 67%	Method is mainly limited to the supervised learning settings
[81]	2019	Proposed a Sequence-to-Sequence conversion model (SCENT) for acoustic feature modeling for voice conversion tasks. The model uses an attention mechanism that implicitly aligns the source and target acoustic feature sequences. At the conversion stages, simultaneous conversion of the acoustic features and duration happens, and finally, the WaveNet vocoder converts the output of the audio model into audio waveforms	Mandarin and CMU Arctic Database	Encoder-Decoder with Attention Mechanism + Vocoder	MCD = 3.802, RMS of F_0 = 41.748, Duration differences of converted and target utterances (DDUR) = 0.260, MOS = 3.66	MCD = 3.212, RMS of F_0 = 9.899, Duration differences of converted and target utterances(DDUR) = 0.147, MOS = 4.01	Model is computation expensive, requires a lot of training data, and does not generalize well to the unseen data
[95]	2019	Proposed non-parallel voice conversion methods using an auxiliary classifier VAE. First, the model uses a fully convolutional network to construct an encoder-decoder network to capture long-term time dependencies of acoustic feature sequences of speech. Also, the network uses information-rhetoric regularization during training to ensure that attribute class labels are preserved during the conversion process	Voice conversion Challenge(VCC) dataset 2018	Variational Auto-encoder(VAE)	MCD = ~ 7.01, MOS = 1.75, ABX test for speaker similarity = 25%	MCD = ~ 6.29, MOS = 4.7, ABX test for speaker similarity = 75%	Method require a lot of training data
[96]	2020	Proposed a data-efficient approach to disentangle speech from the noise using Domain Adversarial Training (DAT) for speaker adaptation & speaker encoding methods	MULTI-SPK & CHiME-4	Encoder-Decoder + RNNs	MOS(naturalness) = 3.30, MOS(similarity) = 2.96, MCD = 5.32, Cosine Similarity = 0.85	MOS(naturalness) = 3.67, MOS(similarity) = 3.72, MCD = 3.66, Cosine Similarity = 0.96	Sensitive to complicated acoustic conditional scenarios, e.g., room reverberations. However, the model also lacks in synthesizing utterances of a target speaker

Table 7 (continued)

Ref	Year	Model description	Dataset	Architectural components	Worst Performance (Evaluation Metrics)	Best Performance (Evaluation Metrics)	Limitations
[97]	2020	Proposed a method named Deep conversion for voice conversion that requires a few samples for parallel data processing. Non-parallel data is used to train the general independent speaker model, while parallel data adapts the generalized model towards finer voice conversion between the source and the target speaker	CMU Arctic dataset	LSTM + Vocoder	MOS = 2.09, Preference test = ~ 20%, MCD = 9.32	MOS = 4.63, Preference test = 75%, MCD = 4.11	Model requires training the module separately, so it is computationally expensive
[98]	2020	Proposed a method named ConvS2S-VC for voice conversion based on an entirely conventional network that provides the flexibility of many-to-many, any-to-many real-time natural conversion of the fundamental frequency (F_0), speaker rate, rhythm, and speaker characteristics of the speech	CMU Arctic Database	Encoder-Decoder	MCD = ~ 7.08, Log F_0 (LFC) = 0.771, LDR = 10.85, MOS(speaker simi.) = 2.9, MOS(sound quality) = 2.5	MCD = ~ 5.78, Log F_0 (LFC) = 0.877, LDR = 0.79, MOS(speaker simi.) = 3.5, MOS(sound quality) = 3.9	Method is limited to English language conversion
[99]	2020	Proposed a method based on Phonetic Posteriograms (PPGs) to remove speaker-dependent information in VC models. During the adversarial training, the output of the encoder is fed to the speaker classifier to differentiate the corresponding speaker. Simultaneously, the encoder is optimized to fool the classifier, enabling the encoder to become more speaker-Independent	VCTK Corpus, LibriSpeech corpus and VoxCeleb2 dataset	ED with Attention mechanism + RNN vocoder	MCD = ~ 9.31, VSS(Voice Similarity Score) = ~ 4.80, MOS = ~ 3.77	MCD = ~ 8.37, VSS(Voice Similarity Score) = ~ 5.02, MOS = ~ 3.86	–
[85]	2020	Proposed a StarGAN framework for voice conversion for the target and source speaker unseen in the training dataset. The incremental training approach represents each speaker ID with an embedding vector for voice conversion	VCTK, CMU-Artic & THCHS30	GANs	MOS(reconstruction) = 1.32, MOS(conversion) = 2.89	MOS(reconstruction) = 3.00, MOS(conversion) = 4.00	Model is limited to the processing of two different timbers for voice conversion

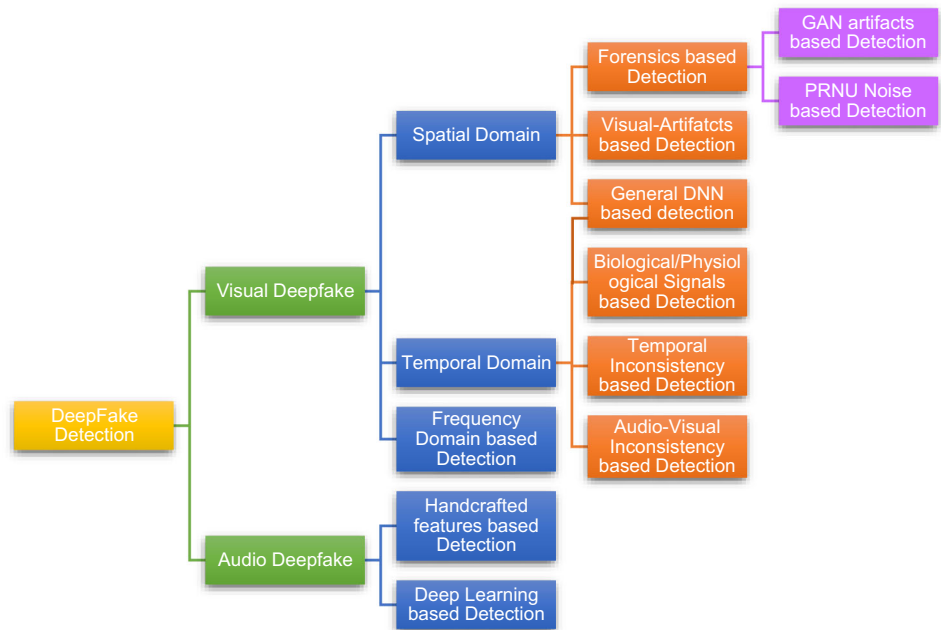
Table 7 (continued)

Ref	Year	Model description	Dataset	Architectural components	Worst Performance (Evaluation Metrics)	Best Performance (Evaluation Metrics)	Limitations
[86]	2020	Proposed an end-to-end voice conversion system based on three concepts: transformer, context preservation and model adaption. The transformer connects all positions, enabling the network to learn long-term dependencies. The context preservation mechanism accelerates the training, while the adaptation mechanism empowers the training to conduct smoothly with limited training data	Mandarin dataset	Transformer Networks + WaveNet vocoder	MOS(naturalness) = ~ 3.84, MOS(similarity) = ~ 4.11	MOS(naturalness) = ~ 4.29, MOS(similarity) = ~ 4.31	Method suffers from low inference and sometimes produces critical errors
[100]	2020	Extend the network of Cycle-GAN to be used for multi-speaker by using conditional cycle-GAN. The proposed network uses a single generator and discriminator conditioned on a pair of source and target identity vectors to steer the network to synthesize the speech for the target speaker	VCC2018 corpus	GANs	MCD = ~ 7.85, MSD = ~ 1.83, MOS(Sound quality) = ~ 3.00, MOS(Similarity test) = 28%	MCD = ~ 7.25, MSD = ~ 1.89, MOS(Sound quality) = 3.92, MOS(Similarity test) = 69%	Method does not work well for the unseen data
[101]	2021	Proposed a novel voice-conversion-based network using the transfer learning mechanism from TTS-based systems. The encoder takes the speech as an input instead of text, while the decoder has similar functionality to TTS systems. However, it is conditioned on speaker embedding to be trained for non-parallel data for any-to-any voice conversion. While training text or speech can be an input to the TTS and VC network, the VC network uses speech as an input during run-time	train-clean-360, VCC2018, AISHELL2	Encoder-decoder with Attention mechanism	MCD = 10.38, RMSE = 0.66, AB preference test = ~ 25%, XAB preference test = ~ 25%, BWS test = 4.44, MOS = 2.39	MCD = 3.31, RMSE = 0.27, AB preference test = ~ 95%, XAB preference test = ~ 90%, BWS test = 43.05, MOS = 4.64	Method's performance is unknown for the compressed and noisy voice samples

Table 7 (continued)

Ref	Year	Model description	Dataset	Architectural components	Worst Performance (Evaluation Metrics)	Best Performance (Evaluation Metrics)	Limitations
[102]	2021	Proposed a model to improve the data efficiency of many-to-many StarGAN-based models for voice conversion. To increase the relative number of speakers with limited training data, use a speaker-encoder to extract the speaker embedding from the target speech and then use StyleGAN2 weight adaptive instance normalization (W-AdaIN) on the convolutional layers	VCTK dataset	GANs + Encoder	Accuracy = 97%, EER = 0.66, AB test(naturalness) = 13.3, XAB(similarity) = 36.1	Accuracy = 97%, EER = 0.66, AB test(naturalness) = 13.3, XAB(similarity) = 36.1	Method cannot generate the voices for unseen samples without requiring a lot of such data to train the network
[103]	2021	Uses a method to convert the voice of songs from one singer to the other. The model takes phonetic posterior grams (PPGs) as input and uses an encoder to encode the PPGs and an additional encoder to compress Mel-spectrograms to obtain acoustic and musical information. Furthermore, it uses a singer confusion module and a mel-regressive representation module to improve the melody and timbre performance	Chinese Corpus	Encoder-decoder + vocoder	MOS(similarity) = ~ 3.42, MOS(naturalness) = ~ 3.64, Normalized cross-relation(NCC) = 0.889	MOS(similarity) = ~ 3.57, MOS(naturalness) = ~ 3.75, Normalized cross-relation(NCC) = 0.902	Model is data-intensive and uses a lot of data to generate good-quality voices
[104]	2021	Extending the framework of the Voice Transformer Network (VTN) for voice conversion to simultaneously learn the mapping among different speakers by extracting the various common latent features. In addition, the identity mapping loss is proposed to ensure the input feature sequences remain the same between the source and the target speaker indices, which eventually improves the model's performance at run-time	CMU Arctic Database	Transformer Net	MCD = 6.77, LFC = 0.717, LDR = 2.34, MOS = ~ 3.5, speaker similarity score = ~ 3.5	MCD = 5.91, LFC = 0.835, LDR = 5.41, MOS = ~ 3.5, speaker similarity score = ~ 3.5	Model is computationally expensive and requires different modules to train independently

Fig. 8 Classification of deepfake detection methods based on feature representation



artifacts. Some other techniques ([113, 114, 115, 116]) also have been proposed, which look for visual clues for deepfake detection.

Temporal domain-based detection *Temporal information* is the sequential info that is relatable, coherent and changes with time. Manipulated artifacts or traces could appear as flickering/jittering. The subsequent sections will discuss various methods that investigated the temporal domain to find such clues. Table 9 gives a brief overview of temporal domain-based detection approaches.

Audio-visual inconsistency based detection For lip-sync methods, inconsistency between the mouth region (visual) and audio is one distinguishing factor for deepfake detection. Agarwal et al. [117] use a CNN to detect inconsistent mouth features, i.e., the shape of the mouth (visemes) is not aligned with spoken words (phenomes) in a manipulated video. They focused on the visemes related to the words M, B and P, in which the mouth almost gets completely closed. However, their method is specific to the videos of Barack Obama. Mittal et al. [118] simultaneously exploited visual & audio modalities and perceived the affective cues to detect any alteration. This exploration between these two modalities uses the Siamese network, where triplet loss is used to measure the similarity. Sometimes, this method cannot detect any manipulation if there is a similarity between these perceived affective cues. Moreover, this method is limited to a single person in a video. Chugh et al. [119] proposed an approach to calculate disharmony score between visual and audio modality and termed it Modality Dissonance

Score(MDS). This dissimilarity score is calculated chunk-wise per video segment. The contrastive loss is employed to calculate such inter-frame modality similarity. However, these methods rarely look into visual consistencies, which could also be faked. It remains unclear whether such methods can be deployed in real-world scenarios where one may encounter any manipulation and undergo different post-processing operations.

Temporal inconsistency based methods (TI) In manipulating video frames, correlation structures between the frames are sometimes destroyed, reflecting in various forms like video flickering or shifting of the facial content [12]. Usually, sequence models like RNNs and their variants get employed to find such inconsistencies between the frames. Hosier et al. [120] use video speed manipulation as a temporal feature for detection, and the encoding used for each frame gives an idea about the number of deleted and added frames. Methods that use frame-level artifacts and temporal features for detection perform better than those focusing on either of the two. Guera et al. [1] propose a temporal-aware CNN-LSTM framework, which exploits the frame-level features and temporal inconsistency between frames. Temporal inconsistency introduced by the auto-encoder (used for face swapping) focuses on the face-swapping process unaware of inconsistency introduced by the process, which results in anomaly serving as crucial evidence for detection.

An optical flow mechanism has been used to estimate the per-pixel motion behavior of the adjacent frames. Amerini et al. [121] used a pre-trained model (trained on RGB images) with an optical flow mechanism to capture the dissimilarity

Table 8 Overview of Spatial domain-based deepfake detection methods proposed from 2019 onwards

Methods	Year	Key features	Dataset media	Img	Vid.	Method types	Local.	Input resolu- tion	Architectural components	Database	Worst performance	Best performance	Limitations
[105]	2020	Disparities in the chrominance components of YCbCr and HSV color spaces, especially in the residual domain	•			Forensics based method	✗	128 × 128	SVM	Real images (CelebA, LFW, and FFHQ) but fake images are generated by different GAN models	Acc. = 51.85% FNR = 41.53% FPR = 96.29% (on mismatched image sources)	Acc. = 99.99% FNR = 0.0% FPR = 0.0% (on mismatched image sources)	Performs poorly for real-world scenarios
[106]	2021	RGB domain and noise feature vectors	•			Forensics based method	✓	640 × 640	CNNs	FaceForensics +	AUC = 65.30%	AUC = 99.99%	Performs poorly for the unseen type of face manipulation methods
[108]	2019	GANs fingerprint persisted across different frequency	•			Forensics(GAN artifact)	✗	128 × 128	CNN	CelebA, LSUN	Accuracy = 97.04% FD Ratio = 6.27	Accuracy = 99.93% FD Ratio = 454.76	Performs poorly on post-processing operations
[163]	2019	Suppressed saturated and under-exposed pixel frequency to amplify GAN traces	•			Forensics(GAN artifact)	✗	–	SVM	Media Forensics Challenge 2018 datasets, CelebA HQ	AUC = 0.61	AUC = 0.92	Does not consider the entire spectrum of GAN-generated images
[164]	2020	Convolutional traces of the image generation process	•			Forensics(GAN artifact)	✗	216 × 216, 256 × 256, 1024 × 1024	KNN, SVM, LDA	Real Images(CelebA), Fake images(AttGAN, GDWCT, StarGAN, StyleGAN, StyleGAN2)	Acc.(KNN) = 71.17% Acc.(SVM) = 76.50% Acc.(LDA) = 76.00%	Acc.(KNN) = 99.61% Acc.(SVM) = 99.81% Acc.(LDA) = 99.61%	Not robust enough to the deepfakes generated using standard image editing(e.g., compression, photometric) operations

Table 8 (continued)

Methods	Year	Key features	Dataset	Method types	Local	Input resolution	Architectural components	Database	Worst performance	Best performance	Limitations
			media								
			Img	Vid.							
[165]	2020	GAN traces	<ul style="list-style-type: none">•	Forensics(GAN artifact)	✗	224×224	CNNs(ResNet)	FaceForensics + & other image are generated by different GANs	mAP = 81.3	mAP = 93.0	Not generalize enough to real-world images
[166]	2021	PRNU features are extracted as spectral and spatial features	<ul style="list-style-type: none">•	Forensics(PRNU)	✗	10×10	SVM	FaceForensics + Deepfake-TIMIT, DeepfakeDetection	Accuracy = 60%	Accuracy = 90%	Does not outperform the state-of-the-art deep learning-based methods
[110]	2019	Warping artifacts of the synthesizing methods	<ul style="list-style-type: none">•	Visual artifact	✓	224×224	CNNs	UADFV & DeepfakeTIMIT	AUC = 93.2%	AUC = 99.9%	Not robust enough for multiple video compression
[112]	2019	Visual inconsistencies of eye color, illumination inconsistency, facial textures & geometry	<ul style="list-style-type: none">•	Visual artifact	✗	–	kNN, LogReg & MLP classifier	CelebA, FaceForensics	AUC(Gen. faces) = 0.704 AUC(Deepfake) = 0.402 AUC(Face2face) = 0.654	AUC(Gen. faces) = 0.852 AUC(Deepfake) = 0.851 AUC(Face2face) = 0.866	Limited to the specific dataset of face images with open eyes and clear teeth
[111]	2020	Blending boundaries of face manipulation	<ul style="list-style-type: none">•	Visual artifact	✓	64×64	CNNs(HRNet)	FaceForensics + Celeb-DF, Deepfake Detection Challenge dataset	AUC = 70.56% AP = 68.99% EER = 32.62% Acc. = 85.69%	AUC = 99.31% AP = 93.34% EER = 8.37% Acc. = 97.73%	Unable to detect the entire synthetic image
[167]	2020	Discrepancies of the tightly bounded face and its context	<ul style="list-style-type: none">•	Visual artifact	✗	299×299	CNNs	FaceForensics + DFDC, Celeb-DF-v2	AUC = 66% Accuracy = 75.19%	AUC = 99.7% Accuracy = 75.49%	Performs badly for the low contrast and Blurry features samples

Table 8 (continued)

Methods	Year	Key features	Dataset media	Method types	Local. Input resolu- tion	Architectural components	Database	Worst performance	Best performance	Limitations
[168]	2021	Subtle texture differences of the image saliency	• Img Vid.	Visual artifact	✗	CNNs	FaceForensics +	Accuracy = 0.9926	Accuracy = 0.999	Computationally expensive and performs poorly for the data outside the training data scope
[169]	2021	Angular symmetrical differences in faces	•	Visual artifact	✗	CNNs	DF-TIMIT, FaceForensics +, DFD, DFDC & Celeb-DF	AUC = 0.552	AUC = 1.00	May not work for the entire synthetic frames with excellent angular symmetry
[170]	2021	Landmark detection over an extended range	•	Visual artifact	✗	Transformer-encoder	Celeb-DF, DFDC, FaceForensics	Accuracy = 74.35%, AUC = 92.34%	Accuracy = 99.2%, AUC = 94.96%	Not generalizable to the unseen type of face manipulations
[171]	2021	Multi-scale texture differences	•	Visual artifact	✗	Encoder-decoder	Ff ++, Celeb-DF, Deeperforensics, DFDC, Celeb-DF	AUC = 0.9579, Acc. = 55.46%	AUC = 0.9999, Acc. = 99.86%	Not generalizable to the unseen data
[172]	2021	Textures dynamics along spatial and temporal dimensions	•	Visual artifact	✗	SVM	FaceForensics +	Accuracy = 72%, AUC = 0.80%	Accuracy = 89.43%, AUC = 0.94%	Unsatisfactory results on the highly compressed videos
[173]	2021	Multi-scale Texture difference using pixel intensity and pixel gradient information	•	Visual artifact	✗	CNNs	Ff ++, DFDC, DeeperForensics1.0, Celeb-DF	Accuracy = 53.43%(cross-database)	Accuracy = 99.99%(cross-database)	Requires to re-train for unknown face manipulation

Table 9 Overview of Temporal domain-based deepfake detection methods proposed from 2019 onwards

Methods	Year	Key features	Dataset media		Method types	Local.	Input resolution	Architectural components	Database	Worst performance	Best performance	Limitations
			Img	Vid.								
[123]	2019	Inconsistent head poses are determined using facial landmark points	•		Physiological Signal	✗	64 × 64	SVM	UADFV, DARPA MediFor GAN	AUC = 0.843	AUC = 0.89	Performance degradation for Blurry images
[174]	2019	Unique trait of facial expression and head pose of a speaker	•		Physiological Signal	✗	–	SVM	Videos of world leaders taken from Youtube	AUC = 0.92	AUC = 1.00	Performance is compromised when the person is looking off the camera
[125]	2020	Heartbeat rhythms of an image identity	•		Biological signal	✗	–	CNNs + LSTM	FaceForensics ++ & DFDC-preview	Accuracy = 0.975	Accuracy = 0.997	Not generalizable to unseen dataset
[126]	2020	Spatial coherence and temporal consistency of biological signals	•		Biological Signal	✗	128 × 128	CNNs + SVM	Ff, Ff ++, Celeb-DF, UADFV & Deep Fakes dataset	Accuracy = 67.61%	Accuracy = 99.39%	Not robust enough for complex and diverse scenarios
[175]	2020	Discriminative lip feature indicates the talking style of an individual	•		Physiological Signal	✗	–	CNNs	GRID & MOBIO	FRR = 2.7 HTER = 6.0	FRR = 0.4 HTER = 0.6	Approach is based on the visual speaker authentication, which cannot be considered a general detection mechanism

Table 9 (continued)

Methods	Year	Key features	Dataset media		Method types	Local.	Input resolution	Architectural components	Database	Worst performance	Best performance	Limitations
			Img	Vid.								
[176]	2021	Semantic inconsistencies in the emotions of human speakers are predicted using the subject face and voice	•		Biological Signal	✗	–	LSTM, RF, XGBoost, LR, kNN	SEMAINE, DFDC	AUC = 0.921, Accuracy = 80.1%	AUC = 1.00, Accuracy = 99.5%	Method is biased to the Caucasian adults with British accents dialects; hence it may not generalize to other cultures
[177]	2021	Inconsistencies in the synthetic eyes and gazes	•		Physiological Signal	✗	–	CNNs	Ff ++, Celeb-DF, Deep Fakes & DeeperForensics	Accuracy = 80.0%	Accuracy = 99.27%	LSTM model's performance decreased with the smaller dataset
[178]	2021	Inconsistencies of the corneal specular highlights between the two eyes	•		Physiological Signal	✗	1024 × 1024	–	FFHQ, StyleGAN2 Images	AUC = 0.94		Method compared the pixel differences without considering the geometry and scene and assumes that images obey a particular portrait setting
[179]	2021	Lip-sync dynamics, where ear movement decoupled from the mouth and jaw movement	•		Physiological Signal	✗	256 × 256	Logistic Regression Model	Videos are downloaded from the Youtube	AUC = 0.77	AUC = 0.97	Method may not work for long hair, large head movement or occluded ears
[121]	2019	Inter-frame dissimilarity	•		Temporal inconsistency	✗	300 × 300	CNNs	FaceForensics ++	Accuracy = 75.41%	Accuracy = 81.61%	Method has reported very few results

Table 9 (continued)

Methods	Year	Key features	Dataset media		Method types	Local.	Input resolution	Architectural components	Database	Worst performance	Best performance	Limitations
			Img	Vid.								
[180]	2019	Temporal inconsistencies exhibit low-level artifacts of the face region	•		Temporal inconsistency	✗	224 × 224	CNN + GRU	FaceForensics ++	Accuracy = 94.35% AUC = 95.93%	Accuracy = 96.9% AUC = 99.59%	Approach is limited to FaceForensics ++ dataset
[181]	2019	Temporal discrepancies across image streams along with the landmark face alignment features	•		Temporal inconsistency	✗	224 × 224	CNNs + GRU	FaceForensics ++	Accuracy = 94.35% AUC = 0.9593	Accuracy = 96.9% AUC = 0.9964	Results are reported for the compression version of FaceForensics ++ frames only
[120]	2020	Video speed manipulation feature	•		Temporal inconsistency	✗	–	SVM	Deepfake detection challenge 2019	Accuracy = 83.1%	Accuracy = 99.1%	Method may not detect the manipulation done within the existing frames of the video
[182]	2020	Structural inconsistencies of the frames and inter-frame temporal discrepancies	•		Temporal inconsistency	✗	256 × 256	CNNs + LSTM	FaceForensics ++ & videos from the Youtube	Accuracy = 93.33%	Accuracy = 95.25%	Method doesn't exploit frame-level inconsistencies. Also, performance degradation for compressed videos

Table 9 (continued)

Methods	Year	Key features	Dataset media		Method types	Local.	Input resolution	Architectural components	Database	Worst performance	Best performance	Limitations
			Img	Vid.								
[122]	2021	Inter-frame temporal structural dissimilarities	•		Temporal inconsistency	✗	224 × 224	CNNs	FaceForensics ++	Accuracy = 71.65%	Accuracy = 98.41%	Approach is limited to the FaceForensics ++ dataset
[183]	2021	Temporal features guided by the attention mechanism	•		Temporal inconsistency	✗	224 × 224	CNNs with Attention mechanism	FF ++ and Celeb-DF	AUC = 0.8278, Accuracy = 87.57%	AUC = 0.998, Accuracy = 99.29%	Scope of improvement for the model's generalization capability and the model is not robust for a different compression level
[184]	2021	Temporal inconsistencies of the frames using Prototype network	•		Temporal inconsistency	✓	–	CNNs	FaceForensics ++, DFD, DeeperForensics, Celeb-DF	AUC = 68.20%, EER = 37.08% (unseen)	AUC = 92.44%, EER = 16.21% (on unseen dataset)	–

Table 9 (continued)

Methods	Year	Key features	Dataset media		Method types	Local.	Input resolution	Architectural components	Database	Worst performance	Best performance	Limitations
			Img	Vid.								
[118]	2020	Disparity between audio (speech), video (face) modality, and perceived affective cues	•		Audio-Visual Inconsistency	✗	–	CNNs	DF-TIMIT, DFDC	AUC = 84.4%	AUC = 96.3%	Approach is limited to a single person per video. Model misclassifies when a Deepfake video does not contain a mismatch between modalities
[117]	2020	Mouth inconsistencies between the shape of the mouth (visemes) with spoken words (phenomes)	•		Audio-Visual Inconsistency	✗	128 × 128	CNNs	Videos created using A2V, T2V-S, T2V-L methods & videos also taken from Youtube	Accuracy = 93.9%	Accuracy = 97%	Method is limited to the MBP phonemes and the videos of Barack Obama
[119]	2020	Disharmony between audio and video modalities	•		Audio-Visual Inconsistency	✓	224 × 224	CNNs	DFDC, DF-TIMIT	AUC = 91.5%	AUC = 97.9%	Does not look into visual inconsistency in the samples, which could also be faked

between frames. However, the methods have reported very few results. Caldelli et al. [122] also proposed optical flow-based CNNs that exploit the motion dissimilarities in the temporal nature of the video sequences using optical flow fields. Again the approach is limited to the specific dataset. These techniques tend to perform better as they are independent of the specific type of manipulation. However, the overall performance improves when these methods are used in conjunction with spatial methods.

Physiological/biological signals based methods (PBS) In deepfake videos, inconsistencies are exhibited either at the physiological level (like inconsistencies in eye blinking patterns, head poses, blending boundaries) or biological level (inconsistencies in a heartbeat), exploited for deepfake video detection. Methods can generate deepfakes with high realism, but they cannot replicate every reasonable behavior, leading to inconsistencies. Such inconsistencies at the physical level may or may not be visible to the eyes; hence, some landmark detector that captures the coordinates of the desired location can be used further for classification. Yang et al. [123] use a landmark detector for 3D head poses to calculate their estimated positions, exploited by the SVM classifier. Their performance degrades to blurry images. Li et al. [124] proposed a method based on eye blinking patterns, which is not well preserved in the synthesized videos. Synthesized videos usually have less frequency of eye blinking, which leads to their detection. The technique can exploit abnormalities in the normal functioning of an organ for deepfake video detection. These signals are preserved neither spatially nor temporally, and different architectures exploit them for detection. Qi et al. [125] use heartbeat rhythms as a clue for detecting deepfake videos. Visual photoplethysmography (PPG) monitors the heartbeat rhythms and captures the abnormalities of the deepfake videos. However, the model does not generalize well to the unseen dataset. Ciftci et al. [126] proposed FakeCatcher methods that use biological signals such as heart rate to exploit the authenticity of a video. They have extracted signals on a pairwise basis, transformed them to a different domain (like frequency, time, etc.), and used this transformation for classification.

Although the methods based on these features perform well on the various datasets, such signals get seriously affected by the video's quality and hence a limited application for detection mechanism based on such signals [12]. [127, 128] some other methods that look for biological methods.

Spatial and/or temporal domain-based detection Few detection methods could leverage both the spatial and the temporal domain or either of the two. However, features fetched along both domains capture the broader range of

manipulation traces which would eventually help in a better detection mechanism. The following section will discuss such detection methods.

General DNN-based detection Instead of focusing on the specific artifacts, some researchers let the network decide which latent features to analyze and learn the mapping accordingly. Deep neural networks drive such methods. Khalid et al. [129] use variational auto-encoder (VAE) to train with the real images, classify them, and treat others as anomalies. The generalization of the unseen dataset has been a more significant issue for such methods. They learn the specific type of manipulation on the data they are trained upon and hence tend to overfit and perform poorly on the other kind of manipulation in the wild. However, few authors have developed a generalizable detector. Xuan et al. [130] also proposed a generalized GAN images forensics detector that preprocesses the images with Gaussian blur and noise operation to enhance the high-frequency pixel noise to allow the CNN model to learn intrinsic discriminative features.

Nowadays, pre-trained models are used heavily to use the learned weights of similar problems to reduce the time complexity of the network. ResNet, XceptionNet, Densenet, AlexNet, and Inception models are recent state-of-the-art models generally used as pre-trained models. Zhou et al. [131] used GoogleNet Inception V3 pre-trained model in one of the branches of the architecture to detect the tampered artifacts evidence and noise inconsistency. Jeon et al. [132] proposed a framework for neural talking head detection using a pre-trained AlexNet model to extract features from a highly unbalanced dataset and then classify them further using a Siamese network-based classifier.

DNN models could also be used for sequence-based problems where data in audio or video frames pass through the model to analyze and learn the intrinsic pattern. Recurrent Neural Networks and their variants LSTM and GRU are generally used for the same purpose. Wu et al. [133] exploited temporal, spatial, and steganalysis features for deepfake video detection. The deep neural network extracts spatial features for the tampering artifacts like irregular shapes, color, etc. Steganalysis features are extracted by putting constraints on the convolution filter for underlying abnormal statistics of the pixels. The temporal inconsistencies are extracted using RNNs. This method beats the current state-of-the-art methods on the FF++ dataset.

DNN-based methods are very good at learning and extracting the intrinsic characteristics of several domains. Such methods tend to overfit the specific datasets they are trained upon, but they lack generalizability to other datasets. Also, the existing methods fail to prove their effectiveness against the adversarial noise attacks [11]. Also, the models do not have the interpretation of why their method proved something fake courtesy of the black-box nature of the model.

Some other methods [134, 135, 136, 137, 138, 139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 154, 155, 156, 157, 158, 159, 130] have extracted discriminative features using DNN either implicitly or explicitly to perform deepfake detection.

Table 10 highlights the various DNNs approaches.

Frequency domain-based detection The frequency domain represents the change of pixel distributions along the different axis. Real images have a frequency distribution; such difference could be revealed in the frequency domain when the generative model does the manipulation. Frank et al. [160] explored the frequency spectrum of the GAN-generated images. They found that images exhibit severe artifacts consistent across different resolution images, caused mainly by the upsampling process of the GANs. Furthermore, the model is robust against image perturbation like blurring, cropping, compression and noise addition. Durall et al. [161] observe the behavior of the real and fake images in the classical frequency domain and use such behavior to be detected by the classifier. However, the model has low accuracy on the low-resolution images. Masi et al. [162] used a two-stream network to exploit frequency domain information in one of the streams using the Laplacian of Gaussian (LoG) operator. The LoG acts as a band-pass filter to suppress the image content and amplify the artifacts. Nevertheless, the method struggles to detect real-world data samples. Table 11 gives an overview of frequency domain-based detection methods.

Summary of the visual deepfake detection methods There has been a similar trend for various visual detection methods. Most methods use the deep learning-based classifier, while traditional machine classifiers like SVM, KNN, and RBF have also been used. AUC and accuracy have been the most preferred evaluation metric for accessing the model's performance. Also, researchers have preferred the public dataset to evaluate their model rather than designing their dataset. Most of the methods are doing the detection only; there is no localization of the affected area, especially in the case of video media. Three key issues continue to challenge the research community: generalizability of the detection mechanism, robustness against post-processing operations and adversarial noise attacks, and interpretability of the detection mechanism. However, few researchers have solved them to a certain extent; there is a long way before these systems can be implemented for real-world scenarios [11].

2.2.2 Deepfake audio detection

Audio deepfake detection is very different from video or image detection. Voice signals in audio are one-dimensional; artifacts would be hard to find in audio but provide sufficient evidence for investigation [199]. Voice is recorded where

much noise is present, and then in such circumstances, it is easy for an attacker to fool the detector by adding the real-world noises [199]. Hence, there is a need for a generalized robust detector that detects every kind of manipulation. In addition, with the increase in authentication systems, people prefer to use automated speaker verification as a biometric system where a voice sample is used to authenticate the identity. However, such systems are vulnerable, as attackers easily get all biometric traits required for spoofing such biometric systems. Apart from using voice biometric spoofing, audio deepfake can also be misused in phishing scams, spreading misinformation, fake evidence, blackmail, and bullying [71]. Such vulnerabilities can only be restricted by developing such detector that detects their use and inform about their misuse. The detection mechanism is divided into handcrafted-based and deep learning-based detection methods based on the feature representation.

Handicraft feature-based methods (HC) Audio features are extracted from the audio signals by the hand-engineered algorithm; after processing them, they are fed into the classifier to learn the classification. Various audio features represent the relevant aspects of the audio signals extracted using other techniques. MFCCs (Mel-frequency cepstral coefficients), Cepstral coefficients, spectrogram, Constant Q Cepstral Coefficients (CQCCs), Linear Predictive Cepstral Coefficients (LPCCs), Spectral Sub-band Centroid Coefficients, and Complex Cepstral Coefficients (CCCs) are a few of the well-known acoustic features used in the different situation [200]. E.g., MFCC represents the perceptual aspect of the speech spectrum, and it is the most commonly used audio feature [200]. The most widely used conventional classifiers are GMM, GMM-UBM, and SVM. Saranya et al. [201] proposed a model that uses reverberation and channel information of a non-voiced segment of audio signal features (MFCC, CQCC, and MFS) and makes a classification using a GMM classifier. Many Replay attack detection methods investigate the frequency spectrum, as replay recording has noise added (in practical scenarios) reflected in the audio signal's frequency spectrum. Witkowski et al. [202] proposed a model that exploits the high-frequency signal of the replayed recording for spoof detection. AlBadawy et al. [203] did a bi-spectral analysis to find the high order spectral correlation introduced by the speech synthesis mechanism.

Deep learning-based methods (DL) Deep learning algorithms' performance is on the rise, and they perform much better than the conventional methods because of the feature abstraction capability, which a deeper model can learn [204]. The GMM classifier only captures first and second-order statistics of the feature vectors, while a deep neural network can also capture linear and nonlinear aspects of the features, giving robustness to various detection methods [204]. Wang

Table 10 Overview of General DNNs based deepfake detection methods proposed from 2019 onwards

Methods	Year	Key features	Dataset media		Local.	Input resolution	Architectural components	Database	Worst performance	Best performance	Limitations
			Img	Vid.							
[185]	2019	Latent space distribution features	•	•	✗	256 × 256, 128 × 128	Auto-encoder-decoder	FaceForensics, CelebA-HQ (real images) & diff. GANs samples	Acc. (synthetic images) = 82.05% Acc. (Inpainting) = 70.62% Acc. (face2/swap) = 72.57%	Acc. (synthetic images) = 100% Acc. (Inpainting) = 99.77% Acc. (face2/swap) = 94.47%	Computationally expensive due to large feature space
[186]	2019	Cross-layer common discriminative features	•		✗	64 × 64	CNNs-DenseNet	CelebA & fake images generated by diff. GANs methods	Precision = 0.909 Recall = 0.865 (for general fake images)	Precision = 0.934 Recall = 0.936 (for general fake images)	Perform poorly for different distribution datasets
[187]	2019	Multi-level abstraction features of the manipulated region	•		✗	64 × 64	CNNs + XGBoost + AdaBoost	SwapMe, FaceSwap	Accuracy = ~ 82.5% AUC = ~ 0.89	AUC = 0.934 Accuracy = ~ 95.1%	Does not detect the entire synthetic image. It also extracts info from RGB channels only, which may always not be present
[132]	2019	Latent discriminative features of pre-trained Siamese network	•		✗	–	CNN-AlexNet	VoxCeleb2	Accuracy = 98.44%, F1 Score = 0.82	Accuracy = 98.84%, F1 Score = 0.98	Methods have shown results to the simpler dataset that are not diverse in nature
[188]	2019	Discriminative features of various spoofs using capsule network	•		✗	128 × 128	Capsule-CNNs	REPLAY-ATTACK dataset, FaceForensics	Acc. (Swap) = 94.47% Acc. (reenact.) = 81.00% Acc. (CG) = 97%	Acc. (Swap) = 99.23% Acc. (reenact.) = 99.37% Acc. (CG) = 100%	Not robust enough against adversarial attacks

Table 10 (continued)

Methods	Year	Key features	Dataset media		Local.	Input resolution	Architectural components	Database	Worst performance	Best performance	Limitations
			Img	Vid.							
[129]	2020	Anomaly scores determined using latent feature characteristics	•	•	✗	100 × 100	Variational Auto-encoder	FaceForensics ++	F1score = 0.707, Acc = 70.70%	F1score = 0.982, Acc = 98.2%	Dependency on the RMSE function to compute the reconstruction score of an image, which may not be a better way
[189]	2020	Visual and temporal dynamics of frames	•		✗	224 × 224	CNNs + RNNs(GRU)	DFDC	Accuracy = 91.88%		Performance needs improvement, and the model dismisses any analysis of audio that could also be faked
[190]	2020	Frame-level features determined by the attention mechanism	•		✗	224 × 224	Attention Map network	Celeb-DF & Youtube videos	Accuracy = 92% AUC = 94%		Method did not consider the audio modality, which could also be faked
[162]	2020	Color and frequency domain features	•		✗	224 × 224	CNNs + LSTM	Ff ++, Celeb-DF, DFDC preview	Accuracy = 86.34% AUC = 73.41%	Accuracy = 96.43% AUC = 99.12%	Not significant performance for practical web-application samples
[133]	2020	Spatial, statistical and temporal features	•		✗	–	CNNs + LSTM	FaceForensics ++ & GAN-based deepfakes	Accuracy = 83.78%	Accuracy = 98.57%	Method requires training the modules separately; hence, it is computationally expensive
[191]	2020	Spatial and temporal semantic features of frames	•		✗	299 × 299	CNNs + LSTM	UADFV, Df-TIMIT, Ff ++, Celeb-DF, DFDC & Youtube videos	Accuracy = 55.91%	Accuracy = 100%	–

Table 10 (continued)

Methods	Year	Key features	Dataset media		Local.	Input resolution	Architectural components	Database	Worst performance	Best performance	Limitations
			Img	Vid.							
[192]	2021	Multi-scale artifacts of manipulated regions	•	✓	–	–	Encoder-decoder	FaceForensics ++, Celeb-DF and UADFV	AUC = 78.68%	AUC = 97.60%	Method doesn't show any robustness against different levels of compression
[193]	2021	Frame-level and temporal inconsistencies of videos	•	✗	256 × 256, 224 × 224	256 × 224 × 224	CNNs	Celeb-DF, FaceForensics + +	AUC = 0.87, Accuracy = 80.05%	AUC = 0.98, Accuracy = 94.64%	Scope of improvement for the robustness of the model against a different level of compression
[183]	2021	Spatiotemporal features use an attention mechanism	•	✗	256 × 256	256 × 256	CNNs(Inception) With Attention mechanism	FF + + & Celeb-DF	AUC = 0.8129 & Accuracy = 85.29%	AUC = 0.9979 & Accuracy = 99.29%	Method does not explain the interpretability of the detection results as to why the samples are declared as fake
[194]	2021	Semantic inconsistencies and noise features	•	✗	128 × 128	128 × 128	CNNs	FaceForensics + +	Accuracy = 85.71%	Accuracy = 90.36%	Method is not generalizable to the unseen type of manipulation
[195]	2021	Multi-domain transfer learning-based features	•	✗	128 × 128	128 × 128	CNNs	FaceForensics ++, DFDC and Celeb-Df	Accuracy = 73.12%(Low-quality videos)	Accuracy = 86.97%(Low-quality videos)	Method is not generalizable to the talking head types of deepfake

Table 10 (continued)

Methods	Year	Key features	Dataset media		Local.	Input resolution	Architectural components	Database	Worst performance	Best performance	Limitations
			Img	Vid.							
[196]	2021	Multi-headed guided attention mechanism to fetch local discriminative texture features	•	✗		380 × 380	CNNs with Attention Mechanism	DFDC, DF, FaceForensics +, and Celeb-DF	AUC = 87.26%, Accuracy = 86.95%	AUC = 99.29%, Accuracy = 97.60%	Method is sensitive to the high compression rate, which curbs the valuable information of the spatial domain
[197]	2021	Temporal characterization over the precise geometric landmark inconsistencies	•	✗		–	RNNs(GRU)	UAADFV, FaceForensics +, Celeb-DF & DeepForensics 1.0	AUC = 55.4%	AUC = 99.9%	Scope of improvement for the model on generalizability parameter. Lack of interpretation for the temporal features results captured by the model
[198]	2021	Intra-frame-level spatial artifacts and the temporal inconsistencies between the frames	•	✗		240 × 240	LSTM	FaceForensics ++, DFD & DFW	F1 score = 87.58%	F1 score = 97.57%	Method does not consider the various level of compression; they performed results only on high-quality videos

Table 11 Overview of frequency-based deepfake detection methods proposed from 2019 onwards

Methods	Year	Key features	Dataset media		Local.	Input resolution	Architectural components	Database	Worst performance	Best performance	Limitations
			Img	Vid.							
[161]	2020	Exhibition of different behavior at higher frequency domain by fake and real samples	•	•	✗	1024 × 1024, 128 × 128	SVM, Logistic Reg. & K-Means classifier	CelebA & FaceForensics +, FF-HQ, 100 K faces project & samples from the thispersondoesnotexist.com	Acc.(SVM) = 58% Acc.(L.R) = 58% Acc.(K-means) = 37%	Acc.(SVM) = 100% Acc.(L.R) = 100% Acc.(K-means) = 94%	Low accuracy on the low-resolution samples
[160]	2020	GAN artifacts exhibited in the frequency domain	•		✗	128 × 128	CNN, kNN & other classifiers	CelebA, LSUN	Accuracy = 69.15%	Accuracy = 99.91%	-

et al. [199] propose a DeepSonar approach that monitors the layer-wise neurons, providing a deep, subtle understanding of AI-synthesized and genuine voice that enables robust detection. Wijethunga et al. [205] use a combination of CNN and RNN models to extract the dynamic acoustic features of the voice samples for detection. However, as various new types of audio detection methods are emerging, the generalizability of the algorithm is still a challenging problem. Chen et al. [206] proposed generalized DNN detection methods; the model uses marginal cosine loss function (LMCL) and frequency masking augmentation to learn robust discriminative features. As discussed earlier, ASV systems are vulnerable to spoofing attacks. There are four kinds of spoofing attacks: speech synthesis (SS), voice conversion (VC), impersonation, and replay attacks. Shim et al. [207] proposed a replay spoofing detection mechanism using the multi-tasking learning approach of DNNs. It is based on the hypothesis that spoofed signal contains replay noise and genuine signal, and internal noise. Yang et al. [208] propose a synthetic speech (SS) detection method by exploiting the high-frequency band of the long-term transform-based features.

Summary of the audio deepfake detection methods Various methods have been proposed for fake audio detection in the past few years. ASV challenge has been devoted to the researcher to develop new approaches to detect counterfeit voices. This challenge also proposed the dataset, which let anyone use and beat up the metrics of the state-of-the-art methods. Table 12 lists the various techniques that have been proposed over the year. The table describes the methods along various features used, type of method, architectural components used, dataset used for evaluating the model, the best performance of the model, and the limitation of the model. There are various insights the listing of methods provides; first, most of the techniques are DNN-based, which shows how such a network learns the latent, intrinsic discriminative features that eventually lead to better results. Second, the ASV challenge provides the dataset and the evaluation metric, EER (and t-DCF also, at times), the typical benchmarks to beat the state-of-the-art methods. Few limitations that visual deepfake detection face is also applied to audio deepfake. Procedures are not generalizable to the unseen type of manipulation, not robust against the various kind of attacks, performance degradation on the low-quality and noisy samples, and lack of interpretability of the detection results. Also, the model is computationally expensive and would need a lot of data for training before it is put for testing. The method needs to tackle these limitations for their deployment in real-world scenarios.

3 Deepfake datasets

With the emergence of a highly sophisticated deepfake algorithm, it becomes apparent to have a dataset that is a good representative of the distribution and enables the methods to leverage their generalizability power. Furthermore, as more and more datasets are evolving, it becomes easy for the detection methods to realize their full potential of detection and arrive at an accurate picture of the performance of their methods. In this section, we will discuss various datasets that have been proposed over the years.

3.1 Image and video deepfake datasets

This category is most sought after as its two visual classes, face swap and face reenactment, carry a huge potential for misuse. Many detection methods have been proposed to standardize the detection benchmark; each brings different manipulations with quantity and heterogeneity. Li et al. [225] have divided the dataset into the most common generations based on release time and synthesis algorithms. UADFV, FaceForensics ++, and Deepfake-TIMIT (DF-TIMIT) datasets belong to the first generations. Google-DFD datasets (later merged to FaceForensics) & Celeb-DF come under the second generation. Finally, Dolhansky et al. [226] proposed a third-generation under which the most recent dataset, DeeperForensics1.0, and DFDC datasets come. Table 13 gives an overview of the various deepfake image and video datasets. Figure 9 presents visual samples for popular visual deepfake datasets.

3.2 Audio deepfake datasets

Fake audio datasets have also unfolded in the past few years with the advent of audio spoof detection methods on biometric systems. In addition, the ASVspoof challenge is conducted, which results in a dataset every third year. Apart from that, various fake synthetic speech TTS datasets have surfaced with increased speech synthesis methods. Table 14 describes such datasets.

4 Architectural components and tools for deepfake

Various deepfake generation and detection frameworks use different components or variations of the below components for their architecture (Fig. 10). Table 15 gives a brief overview of such detection methods.

Today, many open-source tools allow you to create deepfake easily, even without knowing the underlying technology. Table 16 gives a brief description of the popular open-source tools available.

Table 12 Overview of different audio deepfake detection methods from 2019 onwards

Ref	Year	Model and features description	HC	DNN	Architecture components	Dataset	Worst performance	Best performance	Limitation
<i>Hand-crafted features based methods</i>									
[203]	2019	Exploits the high order correlation using bi-spectral analysis to distinguish the synthetic speech from the original human voice	•		LR Classifier	Podcasts Recordings	AUC = 0.88	AUC = 0.99	Performance drops with the addition of noise in the audio samples
[209]	2019	Employed an approach that uses high-order spectral analysis, Gaussianity Statistics & linearity test statistics to capture the artifacts of the speech synthesis generative models	•		–	Baidu cloned Audio dataset	Accuracy = 75%	Accuracy = 100%	Performance remains unclear against the various level of compression and for noisy samples
[210]	2019	Uses Cochlear model to extract the discriminative features of AM and FM speech signal components	•		Cochlea model + GMMs cls	ASV Spoof 2017- 2 replay corpus	EER = 11.30%	EER = 7.32%	Model is computationally complex
[211]	2021	Supervised learning-based model where features are constituted by modeling the speech as an auto-regressive process	•		RF, SVM, RBF SVM	ASV spoof2019	Accuracy = 0.735	Accuracy = 0.741	Performance degradation due to compression

Table 12 (continued)

Ref	Year	Model and features description	HC	DNN	Architecture components	Dataset	Worst performance	Best performance	Limitation
<i>DNN features based methods</i>									
[212]	2019	Uses an attentive Filtering deep neural Network that captures the discriminative feature in the time and frequency domain		•	CNN-ResNet	ASVspoof2017	EER = 8.98%	EER = 8.54%	Method needs a performance improvement
[213]	2019	Uses the DenseNet-LSTM framework that uses hybrid features of the speech segment to make a detection		•	DenseNet-LSTM	ASVspoof2017	EER = 9.56%	EER = 8.84%	Dependency on the background sound of the different record and playback devices; hence model does not perform well in the presence of good-quality hardware
[214]	2019	Proposed an LC-RNN framework that extracts the speech signals features as utterances level embedding used by the back-end classifier		•	CNN-GRU + Classifier (SVM, LDA, PLDA)	ASVspoof2015ASVspoof2017 ASVspoof2019	t-DCF(LA) = 0.1873 EER(LA) = 7.12% t-DCF(PA) = 0.0946 EER(PA) = 3.49%	t-DCF(LA) = 0.1523 EER(LA) = 6.28%, t-DCF(PA) = 0.061, EER(PA) = 2.23%	Method's performance remains unclear in noisy conditions and against various compression levels
[215]	2019	Uses signal-to-noise masks and gated neural network extracts the robust features at utterance level embedding's used by the backend recognizer to spoof audio detection		•	DNNs + GRU	ASVspoof15, ASVspoof17, ASVspoof19	EER = 3.96%((unseen condition)	EER = 2.44%((unseen condition)	Method lacks the interpretation of its processed and detection results
[199]	2020	Monitor the layer-wise neural behavior of DNNs to capture the subtle differences between real and AI-synthesized voices		•	Deep Neural Networks	FoR, Sprocket-VC & MC-TTS	Acc. = 0.981, AUC = 0.982 Score = 0.982 AP = 0.976, EER = 0.021	F1 = 0.982, F1 = 0.982 AP = 0.976, EER = 0.021	Not robust enough against the adversarial attacks and performance degradation for real-world noise samples

Table 12 (continued)

Ref	Year	Model and features description	HC	DNN	Architecture components	Dataset	Worst performance	Best performance	Limitation
[205]	2020	Uses DNN to extract discriminative dynamic acoustic features of voice for classification		•	CNNs + RNNs	For Dataset & AMI corpus	Accuracy = 88%	Accuracy = 89%	Limited to the speaker's utterances one word at a time. Performance degradation for noisy samples
[206]	2020	Uses a Marginal Cosine Loss Function & frequency mask augmentation to learn the robust features through DNNs		•	CNNs	ASVspoof 2019	EER = 4.04% t-DCF = 0.109	EER = 1.26% t-DCF = 0.052	Method not robust enough for noisy samples
[216]	2020	Uses an attention-based framework that extracts LBF features from the selected segments of the high-frequency domain		•	DenseNet BiLSTM Net	BTAS2016, ASVspoof2017	EER = 6.43%	EER = 0.53%	Method requires background sound of different records and playback devices and hence not diverse
[208]	2020	Exploited the high-frequency band info of the long-term transform-based features for synthetic speech detection		•	DNNs	ASVspoof2015 ASVspoof2019	EER = 0.345%	EER = 0.090%	Method does not perform well for real-world scenarios
[217]	2020	Employed a model that exploits the characteristics distribution of the genuine speech and uses it as a feature for classification		•	CNNs + Transformer	ASVspoof2019	EER = 4.07%, t-DCF = 0.102		Approach does not detect the replay attack detection

Table 12 (continued)

Ref	Year	Model and features description	HC	DNN	Architecture components	Dataset	Worst performance	Best performance	Limitation
[218]	2020	Uses an approach where dense connectivity strengthens the propagation of features and ensures the maximum flow of information through dense layers		•	CNNs	ASVspoof19	t-DCF = 0.1853, EER = 8.99%(LA)	t-DCF = 0.0469, EER = 1.98%(LA)	Method does not solve the issue of generalization and robustness against different levels of processing
[219]	2020	Uses a parallel network fetches the RDF features in one branch using CQT, BLSTM, DCT network, while the other network extracts CQCCE acoustic features, fused for classification		•	DNNs + LSTM	ASVspoof2017-V2	EER = 15.08%	EER = 8.89%	Method's robustness against the different kinds of processing remains unclear, and also, it is not generalizable to the unseen samples
[220]	2020	Uses a capsule network modifying the dynamic routing algorithm to focus on the local artifacts to yield better results on the unseen fake audio attacks		•	CNNs	ASVspoof2019	t-DCF = 0.0982 EER = 3.19%(logical access)		Method lacks the interpretability of detection results as to what features the capsule network uses for detection
[221]	2021	Uses a method that extracts the temporal waveform features. It uses the since convolution and squeeze excitation module, which implements the module's recalibration, and captures the interdependencies between the audio channels, enhancing the good bands for detection		•	CNNs	ASVspoof19	EER = 16.39%	EER = 7.23%	Scope of improvement for the generalizability of the algorithm

Table 12 (continued)

Ref	Year	Model and features description	HC	DNN	Architecture components	Dataset	Worst performance	Best performance	Limitation
[222]	2021	Explored the method that extracts the intrinsic deep learning features using CNNs from a raw wave, then CQCC features extracted from the same raw wave		•	LSTM + CNNs	BTAS2016 & ASVspoof2017	EER = 7.73%	EER = 0.79%	Performance is limited to the dataset on which it is trained
[223]	2021	Uses a capsule network that uses different input features to learn various artifacts of spoofed voices that assume to bear the similarity with the spatial information		•	CNNs	ASVspoof2019	t-DCF = 0.1561 EER = 6.32%	t-DCF = 0.1198 EER = 4.93%	Not generalizable to the unknown spoofing attacks
[224]	2021	Modified the ResNet block architecture to capture multi-scale features. It splits the features map within a block into different channel groups and designs a residual-like connection across such groups to enable capturing multi-scale features		•	CNNs	ASVspoof2019	t-DCF = 0.0743 EER = 2.502%(LA)	t-DCF = 0.0452 EER = 1.892%(LA)	Scope of improvement for better generalizability for ASV anti-spoofing tasks

Table 13 Image and Video deepfake datasets

Dataset	Year	Media	Manipulation technique	No. of original samples	No. of manipulated samples	Resolution	Format
Swap Me & FaceSwap dataset [131]	2017	Image	FaceSwap	2300	1005		JPEG
UADFV [123]	2018	Image	FakeApp [14]	49	49	294 × 500	–
Deepfake-TIMIT [227]	2018	Video	FaceSwap-GAN [13]	–	620	64 × 64(LQ), 128 × 128(HQ)	JPG
FaceForensics [228]	2018	Video	Face2face [26] approach	500,000 frames edited from 1004 videos		Videos at least 480p and at most 1080p	Frame rate @ 30fps, either raw with ‘DIB’ codec or compressed with H.264
CelebA-HQ [67]	2018	Image	ProGAN		3000	1024 × 1024	JPEG
Fake Face in the wild(FFW) [229]	2018	Image	CGI, FakeApp, Face Swap	53,000 images from the 150 videos		At least 480p	
FaceForensics + + [16]	2019	Image	Face2Face, FaceSwap, DeepFakes, Neural Textures	More than 1.8 million images from 4000 fake videos		480p, 720p, 1080p	- H.264, CRF = 0,23,40
Deepfake Detection Challenge(DFDC) Dataset [226]	2019	Video	FaceSwap	23,654	104,500	1080 × 1920	H.264
DFDC Preview [230]	2019	Video	Faceswap	1131	4113	1080 × 1920	H.264
Google DFD [231]	2019	Video	Faceswap	363	3068	1080 × 1920	-
Celeb-DF [225]	2020	Video	DeepFake algorithm	590	5639	256 × 256	MPEG4.0
DeeperForensics 1.0 [232]	2020	Video	FaceSwap	50,000	10,000	1920 × 1080	-
WildDeepfake [233]	2020	Face sequences	Faceswap	3805	3509	-	-
Vox-Deepfake [234]	2020	Video	Faceswap	1,045,786	1,125,429	224 × 224	-
Deepfake MNIST + [235]	2021	Animated Video	Siarohin’s framework	10,000	10,000	256 × 256	Raw & H.264 in c23 & c4

5 Current limitations and open challenges

Since the advent of this technology, there has been significant progress; but still, there are open challenges that need to be addressed to make the best use of the technology for beneficial purposes.

5.1 Generation methods

The quality of deepfake samples has evolved with the sophistication of the generative methods; however, several concerns

of these methods would remain the challenge to be foreseen. Various limitations and challenges which may need to look to move forward are as follows:

- Generalizability:** Deepfake generative models are data-driven; to generate the specific identity samples, we need similar identity training data to train the network [4]. It constrains the model to develop the particular sample type, limiting their generalizability. Moreover, it is easier to get enough training data for driving samples than specific examples [4]. In addition, to model each separate identity,







Dataset	Data Samples
Deepfake-TIMIT [227]	
Faceforensics [228]	
Celeb-DF [225]	
Faceforensics++ [16]	
Deepfake Detection Challenge(DFDC) Dataset [226]	
DeeperForensics 1.0 [229]	

Fig. 9 Data samples of popular visual deepfake dataset

we need to train the model again, which would be computationally expensive, making it difficult for the system to deploy anywhere. Therefore, extensive efforts have been put into finding ways to model the methods that need less training data and work on multiple unseen identities.

- **Driver's content dependency:** For face reenactment, content is derived from the driving samples, usually with a frontal pose, constraining the flexibility of pose variations. Also, with the face-swap techniques, the frontal face is derived from the lookalike identity, and finding a good match or similar personality is not possible every time [4]. This limits the flexibility of the generation methods that end with a static performance. Therefore, designing the target identity with expected expression and personality would remain a challenge to be solved.
- **Paired training:** The trained, supervised model generates the desired output based on similar input data. It requires data pairing, which becomes laborious when it is done for every specific identity. The process becomes more complicated when the model is designed for multiple identities for multiple tasks. However, many existing methods resolve this issue by using encoding in the ED network or CycleGAN.

- **Spatial inconsistency and temporal coherence:** Visual-spatial inconsistencies appear in various forms like inconsistent facial features due to interpolation (facial reenactment or lip-sync methods) or blending boundaries (face-swap), which is challenging for few techniques. It also happens due to irregular illumination or abrupt change in lighting conditions of the environment. Another concern is temporal coherence in the form of flickering or jittery among frames of the videos. It usually happens due to considering the individual frames without considering the relationship among them [4]. To overcome these issues, some methods provide this context either to the GANs, employing temporal coherence loss, or using RNNs to consider the temporal relation between the frames or combining both of them [4].
- **Identity leakage:** During the face reenactment or body-puppetry task, the partial driver's identity is transferred to the target identity resulting in inconsistencies in the facial features. It happens when the training is done on a single identity or the network is trained on multiple identities; the data pairing is done with the same identity [4].
- **Heavy reliance on the massive dataset for audio deepfake generation:** To produce the high-quality synthetic speech

Table 14 Audio deepfake datasets

Dataset	Year	Techniques/tools used for generation	No. of Manipulated Samples	No. of Original Samples	Language	Format
CMU Arctic Speech Database [236]	2004	Recorded	1200 phonetically utterances		English	Wave
LibriSpeech [237]	2015	LibriVox's API & Sequitur G2P toolkit	1000 h of Audio book recording		English speech	FLAC(Free Lossless Audio Codec)
ASVSpooof2015 [238]	2016	Audio spoofed by Voice conversion(VC) & Speech Synthesis(SS) algorithms	16,651(Genuine)	246,500(Spoofed)	–	RIFF/Wave
LJ Speech Dataset [239]	2017	Recorded	13,100 short clips of a single speaker		–	Wave
ASVSpooof2017 [240]	2018	Recorded	3565(Bona fide)	14,465(Replay)	English	–
VoxCeleb2 [241]	2018	Two-stream CNNs	1,128,246 number of utterances from 6112 speakers		–	–
CSTR VCTK Corpus [242]	2019	Recorded	110 English Accent		English	–
Fake-or-Real(FoR) Dataset [243]	2019	Deep voice 3, WaveNet, Amazon AWS Polly, Baidu TTS, Microsoft Azure TTS, Google Traditional & Cloud TTS,	111,000 real utterances	87,000 synthetic utterances	English	MP3/WAV
VoxCeleb1 [244]	2019	Two-stream CNNs	153,516 number of utterances from 1251 speakers		–	–
M-AILabs Dataset [245]	2019	Recorded	Around 1000 h recording		German, English, Spanish, Italian, Ukrainian, Russian, French, Polish	Wave
ASVSpooof2019 [246]	2020	Tacotron2 & WaveNet	122,157 Samples		–	Mp3

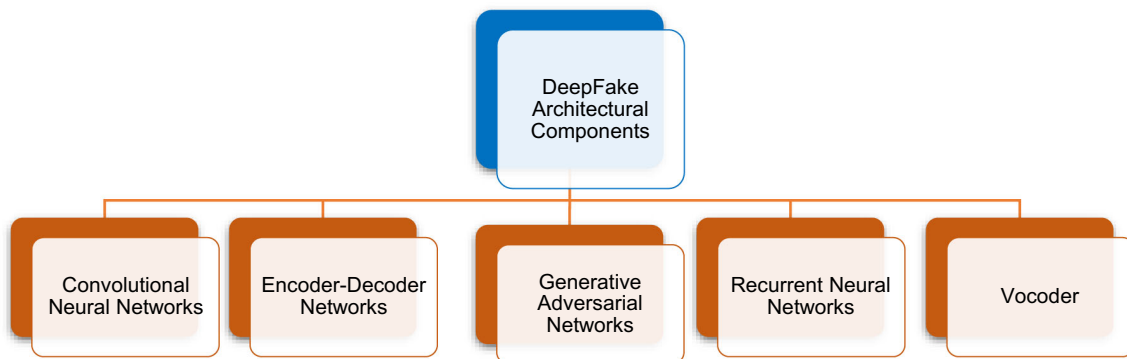
**Fig. 10** Deepfakes Architectural components

Table 15 Different Architecture components of deepfake methods

Components	Descriptions
Convolutional Neural Networks(CNNs)	CNNs extract the visual features from the images or video frames in an abstract feature map using convolutional kernels. Pooling layers are used for dimensionality reduction, while upsampling layers increase it. Non-activation functions like ReLu introduce the nonlinearity that allows learning of nonlinear discriminative features. It is used frequently to analyze the distribution of pixels and extract visual data features using a feature map, which can be used further to manipulate the distribution or find abnormalities for detection
Encoder-Decoder(ED)	Encoder-decoder network is a particular class of neural network where the encoder network learns the mapping of compressed latent features to its distribution. In contrast, a decoder network learns the mapping of compact features to its original input. If the encoder-decoder network is symmetrical and the network regenerates back the original image, then the network is called auto-encoder [4]. Another variant of ED is Variational Auto-encoder (VAE), where the network learns the posterior distribution of input. However, it is better at disentangling the various latent features, which ease the process of modification and interpolation [4]. ED provides the latent space vectors, which different techniques exploit and use to change the attribute in a specific direction. It has also been exploited to identify the distribution pattern of real and fake samples
Generative Adversarial Networks(GANs)	Google researcher Ian J. Goodfellow proposed GANs in 2014 [247]. GANs are composed of two neural networks: Generator (G) and Discriminator (D). Generator generates the fake samples from the random noise distribution while the Discriminator D identifies the artificially generated samples. Both are trained in an adversarial manner with the aim that generator produces more realistic samples while discriminator D gets better at identifying them. Visual deepfake has become so realistic because of the arms race between generator and discriminator that challenge each other to improve the distribution. It goes on and on until we have such real samples that even discriminators have difficulty distinguishing them
Recurrent Neural Networks(RNNs)	RNNs are also a type of neural network that can learn the long-term dependencies of the data; hence, they are good at handling variable length and sequential data. Therefore, it is used to process temporal data like videos and audio. Nowadays, a more advanced form of RNNs is used: Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU). They are very good at learning the dependencies and retaining the order of features to be used for detection or manipulation problems
Vocoder	A voice encoder captures the characteristics of an audio signal that can convert one form of acoustic feature to various other formats. It is used as a component in the varied architecture of speech synthesis, which converts acoustic features to the waveform. Different deepfake techniques are used to extract the acoustic features and give them for the final conversion of acoustic features to the waveform

or audio waveforms, the current model employs deep neural networks that require a lot of training data to learn the speaker's characteristics. Acquiring such extensive data for each speaker is a difficult task, and in addition, it requires retraining the model again for a new speaker, which is a highly computational task. Moreover, few TTS systems require manual annotation for the dataset, which is an uphill task. A few methods use the pre-existing model or employ few-shot learning methods to mitigate this issue, which is a good step in this direction, but there is still a lot to be done.

- **Lack of realism for synthetic voice:** The voice quality is getting better courtesy of the advanced model, but still, there is enough scope for improvement. Artificial voice lacks pauses, accent, tone variation, emotions, and the natural pace of the intended speaker. Mainly, voice-cloning

methods cannot produce a clean synthetic voice and contain the background noise, which can be taken care of by using a clean voice dataset.

- **The realism of real-time deepfakes:** Various methods [26, 256] have generated real-time deepfakes, but they are far from realistic. The realism of the content would be beneficial for imparting education in no time. However, the evolution of these techniques can invite various phishing attacks, which is again a challenge that needs focus.

5.2 Detection methods

Since the inception of deep learning frameworks, the performance of the classification algorithms has gone to the next level; deepfake detection methods are too leveraging their capability for advancing the purpose. Apart from the deep learning-based detection methods, hand-engineered feature-based methods also perform well. Despite all these advances, some serious concern needs to be taken care of to stop this

Table 16 Popular open-source tools available for deepfake generation

Tools	Brief Description	Platform	Manipulation Types	Media	Scale of dataset used
FakeApp [14]	Reddit users developed FakeApp software, and its algorithm has progressed ever since	Desktop Application	FaceSwap	Videos	A single sample of source and target video is required
DeepFaceLab [248]	Pioneer software is famous for manipulating lips and replacing faces and heads. The authors claim to have an accuracy of up to 95%	Open-source implementation	FaceSwap	Videos	A single sample of source and target video is required
Deepfakes web [249]	Online web software has a powerful GPU on the cloud that creates video/image in around 4 h. High-quality deepfakes have a visible watermark to state that it is not real	Web-based Application	FaceSwap	Videos/Images	A single sample of source and target video is required
FaceApp [250]	The smartphone app creates deepfake images using many AI filters and effects, and it garnered attention due to trans-change features	Mobile application	Facial Attribute manipulation	Images	A single sample is required
Faceswap [13]	The open-source tool with lots of functionality allows (even) saving your model's training. It runs on all platforms and requires a high-capability processor GPU to run it	Open-Source Implementation	FaceSwap	Videos	A single sample of source and target video is required
Zao [251]	Smartphones app is mainly used for swapping faces with a celebrity in seconds. Therefore, it works best on Chinese faces	Mobile App	FaceSwap	Videos/Images	A single sample is required
MachineTube [252]	Web-based software allows creating deepfake video/image free. However, it is a bit slow; requires high-end computing GPUs	Web-based App	FaceSwap	Videos/Images	A single sample is required
Doublicat [253]	The smartphone app allows the creation of images, memes, and GIFs quickly. In addition, it makes use of RefaceAI to smoothen the outcomes	Mobile App	FaceSwap	Videos/Images	A single sample is required
Resemble [254]	A text-to-speech (TTS) deepfake software allows cloning, voice modulation, and intonation with an emotion-embedded generated voice	Web-based App	Text-to-speech Synthesis	Audio	Around 5 min clip of audio is required
Wav2Lip [255]	A lip-sync software allows you to sync the mouth region according to the arbitrary audio	Open-source Implementation	Lip-Sync	Videos	A single sample of video and audio files is required

technology's misuse. Following are some concerns that may need to be addressed:

- **Generalization:** Existing deepfake detection methods perform well on the seen dataset, but their performance degrades on the unseen dataset. This concern affects the generality of the algorithms, which makes them unfit for their deployment in real-world scenarios [11]. Moreover, its absence gives the upper hand to the anti-social elements to misuse the technology at their whims and fancies. That's why generalization is one of the most crucial indicators of the performance of the methods [4].
- **Lack of interpretability of methods:** Most detection methods use a neural network with an inherent problem of lack of explainability for their results due to the black-box nature [12]. For real-world scenarios, human-understandable justification is required for any forensic methods. E.g., suppose any deepfake detection tool is deployed in the courtroom to detect the evidence. In that case, reason or explanation may be required for various aspects of it being a deepfake. Therefore, detection methods need to focus on the explainability of the detection results, which undoubtedly remain a concern to look at in the future [11].
- **Scarcity of a large quality dataset:** Dataset plays a pivotal role in the detection methods, as it allows learning the different sets of features required to identify various artifacts [10]. Other proposed datasets suffer from problems due to which detection methods suffer. Some of the noticeable issues:
 - Small in size and not representative of various kinds of manipulation.
 - Inconsistent and blurriness of facial features.
 - Flickering or jittery of video frames.
 - Uneven illumination in images.
 - Lack of noisy audio dataset (to represent real work scenarios).
 - Lack of occluding objects in images.
 - Low-quality images/video frames.
- **Performance degradation in real-world scenarios:** Existing methods tend to perform well in the controlled environment, where we have a dataset that hardly represents real-world scenarios. For real-world data, which contains various noises and manipulation, their performance degrades as the detection methods are designed to identify specific types of artifacts. In addition, real-world data are multi-facet data where subjects are manipulated in various ways, and samples detected as not manipulated do not mean that they have not been changed in any form. Therefore, the binary classification may not be an ideal way to go about it; a multi-class category could be more suitable that tells the varied degree of manipulation [10].

- **Lack of manipulation localization methods:** Various methods can detect the manipulation but cannot localize it. Localization methods, in a way, guarantee that the detection methods have learned the correct features required for detection. In addition, localization may explain the technique used to produce the deepfake, which may be helpful for forensics investigation.

6 Conclusion and future work

There is an arms race between attacker and defender; this race challenges each other to get the best out of them and develop more sophisticated and optimized approaches. However, not all generated deepfakes are malicious. As we saw earlier, deepfakes can be used for creative purposes like education, innovation, entertainment, etc.; it raises the alarm when used for malicious which can afflict pain, exploit and harm individuals, and become a threat to society and nation. This paper comprehensively discussed the application of deepfakes, various types of deepfakes, their generation and detection methods, datasets for multiple media, limitations and open challenges of deepfakes generation and detection methods.

Future holds the promise for this technology to be leveraged fully. Fabricated content would become more and more realistic, which may easily convince naked eyes. Detection tools would be deployed on various platforms to prove the authenticity of the media. Based on papers within this survey, the research community has remarked on the future challenges and work as follows:

- Existing methods primarily focus on the frontal face, face swapping or face reenactment. Future deepfake methods could potentially manipulate the upper or entire body.
- Real-time deepfake generation is expected to be seen more frequently in the future. This deepfake generation technology matures that would require fewer data and generates more believable deepfakes in less time.
- We may expect that deepfake will prevail in other domains in the future. Deep learning approaches generate fingerprints of an individual more convincingly, and deepfakes would be used to create news articles and tweets more frequently. Deepfake can be applied to financial accounts more often to evade detection [4].
- The current trend of deepfake has been focused mainly on generating visual content; in the future, we could see more and more high-quality fabricated audios that resemble human voices in terms of accent, tone variation, emotions, etc.
- We could expect to see human robots speaking in human-realistic voices in the future. These models could be

identity agnostic and trained using a few samples of human voice data. More and more TTS systems would become part of daily chores, and voice impersonation would be easy to obtain.

As the deepfake generation technique continues to evolve in the future, it tends to challenge the detection techniques with the same level of sophistication, which would lead to, some highly optimized, more robust, and generalized approaches. The following may be the future work that would majorly improve the limitations of the existing detection methods:

- Multi-tasking learning strategies may be implemented in future models [12]. Multi-tasking allows doing more than one task at a time, like detection and localization. Each task complements the other, as localization allows the detection procedure to learn the correct artifacts, eventually leading to even a better localization.
- The current problem for the detection methods is to generalize well on the unseen data due to the different distribution of the dataset. A Triplet training strategy could reduce the distance between the identical distributions and widen the gaps between the samples of the other distributions, which subsequently helps in the classification tasks [12].
- Existing detection methods focus on finding the drawbacks of deepfake generation methods and use them as classification artifacts. Such information may not be available in future deepfakes, especially in an adversarial environment where attackers intentionally try not to reveal the artifacts. Futuristic models need to boost approaches' robustness, generalizability, and scalability.
- As this technology grows, social media platforms will be exploited by this technology. There will be a need to integrate deepfake detection methods on such platforms to deal with the widespread malicious effect of this technology. The legal requirement could be framed to enforce this requirement effectively [257].
- The future trend would be employing a unified detection framework that may work on finding multiple forgeries for video media. The framework could use parallel branches to detect the visual content, while the other unit could detect the fake audio. Such a framework could effectively deal with the synthetic content in any form [10].
- Existing datasets are not robust enough to consider various deepfake generation techniques. We have seen fewer data samples based on lip-sync approaches or entire image fabrication. The success of detection approaches is directly proportional to the robustness and variedness of the datasets [10]. Future work could include the proposal of a more varied and robust dataset.
- Deepfake detection methods could prove the evidence in police investigations and court cases. In the future, digital media forensics will be collaborating with the investigation to examine and analyze the prosecutor's proofs [257].

Declarations

Conflict of interest There is no conflict of interest

References

1. Güera D, Delp EJ (2018) Deepfake video detection using recurrent neural networks. In: 2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS), Auckland
2. Strickland E (2019) Facebook AI launches its deepfake detection challenge. In: IEEE, December 2019. <https://spectrum.ieee.org/facebook-ai-launches-its-deepfake-detection-challenge>
3. Chesney R, Citron DK (2018) Deep fakes: a looming challenge for privacy, democracy, and national security, 68
4. Mirsky Y, Lee W (2021) The creation and detection of deepfakes: a survey. *ACM Comput Surv* 54(1):1–41
5. Jaiman A (2020) Positive uses of deepfakes, towards data science, 15 Aug 2020. <https://towardsdatascience.com/positive-use-cases-of-deepfakes-49f510056387>. Accessed 11 April 2021
6. Damiani J (2019) A voice deepfake was used to scam a CEO Out Of \$243,000, Forbes, 3 September 2019. <https://www.forbes.com/sites/jessedamiani/2019/09/03/a-voice-deepfake-was-used-to-scam-a-ceo-out-of-243000/?sh=70583a482241>. Accessed 10 July 2021
7. Jaiman A (2020) Deepfakes harms and threat modeling, 19 Aug 2020. <https://towardsdatascience.com/deepfakes-harms-and-threat-modeling-c09cbe0b7883>. Accessed 14 April 2021
8. Rizzotto L (2019) Deepfake ads, 4 Dec 2019. <https://medium.com/futurepi/why-deepfakes-will-change-advertising-forever-2949ec3f87ee>. Accessed 18 April 2021
9. Tolosana R, Vera-Rodriguez R, Fierrez J, Morales A, Ortega-Garcia J (2020) Deepfakes and beyond: a survey of face manipulation and fake detection. *Inf Fusion* 64:131–148
10. Masood M, Nawaz M, Malik KM, Javed A, Irtaza A (2021) Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward, <http://arxiv.org/abs/2103.00484v1>
11. Juefei-Xu F, Wang R, Huang Y, Guo Q, Ma L, Liu Y (2021) Countering malicious deepfakes: survey, battleground, and horizon. In: <http://arxiv.org/abs/2103.00218v1>
12. Yu P, Xia Z, Fei J, Lu Y (2021) A survey on deepfake video detection. *IET Biometrics* 10(6):607–624
13. Faceswap, <https://faceswap.dev/>. Accessed 6 April 2021
14. FakeApp, <https://www.malavida.com/en/soft/fakeapp/>. Accessed 6 April 2021
15. deepfakes/Faceswap, github, 2016. <https://github.com/deepfakes/faceswap>
16. Rössler A, Cozzolino D, Verdoliva L, Riess C, Thies J, Niessner M (2019) FaceForensics++: learning to detect manipulated facial images. In: IEEE/CVF International Conference on Computer Vision (ICCV), Seoul
17. Dale K, Sunkavalli K, Johnson MK, Vlasic D, Matusik W, Pfister H (2011) Video face replacement. *ACM Trans Gr* 30(6):1–10
18. Li L, Bao J, Yang H, Chen D, Wen F (2020) Advancing high fidelity identity swapping for forgery detection. In:

- IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle
19. Nirkin Y, Keller Y, Hassner T (2019) FSGAN: subject agnostic face swapping and reenactment. In: IEEE/CVF International Conference on Computer Vision (ICCV), Seoul
20. Chen R, Chen X, Ni B, Ge Y (2020) SimSwap: an efficient framework for high fidelity face swapping. In: Proceedings of the 28th ACM International Conference on Multimedia, Seattle
21. Zhu Y, Li Q, Wang J, Xu C, Sun Z (2021) One shot face swapping on megapixels. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville
22. Zhang L, Yang H, Qiu T, Li L (2021) AP-GAN: improving attribute preservation in video face swapping. *IEEE Trans Circuits Syst Video Technol (Early Access)* 32(4):2226–2237
23. Peng B, Fan H, Wang W, Dong J, Lyu S (2021) A unified framework for high fidelity face swap and expression reenactment. *IEEE Trans Circuits Syst Video Technol (Early Access)* 32(6):3673–3684
24. Cao M, Huang H, Wang H, Wang X, Shen L, Wang S, Bao L, Li Z, Luo J (2021) UniFaceGAN: a unified framework for temporally consistent facial video editing. *IEEE Trans Image Process* 30:6107–6116
25. Chan C, Ginosar S, Zhou T, Efros A (2019) Everybody dance now. In: IEEE/CVF International Conference on Computer Vision (ICCV), Seoul
26. Thies J, Zollhöfer M, Stamminger M, Theobalt C, Nießner M (2016) Face2Face: real-time face capture and reenactment of RGB videos. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas
27. Thies J, Zollhöfer M, Nießner M (2019) Deferred neural rendering: image synthesis using neural textures. *ACM Trans Gr* 38(4):66
28. Liu L, Xu W, Zollhöfer M, Kim H, Bernard F, Habermann M, Wang W, Theobalt C (2019) Neural rendering and reenactment of human actor videos. *ACM Trans Gr* 38(5):1–14
29. Christos Doukas M, Koujan MR, Sharmanska V, Roussos A, Zafeiriou S (2021) Head2Head++: deep facial attributes re-targeting. *IEEE Trans Biometrics Behav Identit Sci* 3(1):31–43
30. Zakharov E, Shysheya A, Burkov E, Lempitsky V (2019) Few-shot adversarial learning of realistic neural talking head models. In: IEEE/CVF International Conference on Computer Vision (ICCV), Seoul
31. Wang T-C, Liu M-Y, Tao A, Liu G, Kautz J, Catanzaro B (2019) Few-shot video-to-video synthesis. In: Advances in Neural Information Processing Systems (NeurIPS), Vancouver
32. Gafni O, Ashual O, Wolf L (2021) Single-shot freestyle dance reenactment. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville
33. Zhang J, Zeng X, Pan Y, Liu Y, Ding Y, Fan C (2019) FaceSwap-Net: landmark guided many-to-many face reenactment. <http://arxiv.org/abs/1905.11805v1>
34. Zhang Y, Zhang S, He Y, Li C, Loy CC, Liu Z (2019) One-shot face reenactment. <http://arxiv.org/abs/1908.03251v1>
35. Gu K, Zhou Y, Huang T (2020) FLNet: landmark driven fetching and learning network for faithful talking facial animation synthesis. In: Proceedings of the AAAI Conference on Artificial Intelligence, Hilton New York Midtown
36. Lee J, Ramanan D, Girdhar R (2020) MetaPix: few-shot video re-targeting. In: International conference on learning representations
37. Sanchez E, Valstar M (2020) A recurrent cycle consistency loss for progressive face-to-face synthesis. In: IEEE international conference on automatic face and gesture recognition, Buenos Aires
38. Tripathy S, Kannala J, Rahtu E (2021) FACEGAN: facial attribute controllable rEnactment GAN. In: IEEE winter conference on applications of computer vision (WACV), Waikoloa
39. Lee C-H, Liu Z, Wu L, Luo P (2020) MaskGAN: towards diverse and interactive facial image manipulation. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR), Seattle
40. Zhu Z, Huang T, Shi B, Yu M, Wang B, Bai X (2019) Progressive pose attention transfer for person image generation. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR), Long Beach
41. Aberman K, Shi M, Liao J, Lischinski D, Cohen-Or D, Chen B (2019) Deep video-based performance cloning. In: European association for computer graphics, Genova
42. Zhou Y, Wang Z, Fang C, Bui T, Berg TL (2019) Dance dance generation: motion transfer for internet videos. In: IEEE/CVF international conference on computer vision workshop (ICCVW), Seoul
43. Tripathy S, Kannala J, Rahtu E (2020) ICface: interpretable and controllable face reenactment using GANs. In: IEEE Winter Conference on Applications of Computer Vision (WACV), Snowmass
44. Zablotskaia P, Siarohin A, Zhao B, Sigal L (2019) DwNet: dense warp-based network for pose-guided human video generation. In: British Machine Vision Conference (BMVC), Cardiff
45. Suwajanakorn S, Seitz SM, Kemelmacher-Shlizerman I (2017) Synthesizing Obama: learning lip sync from audio. *ACM Trans Gr* 36(4):1–14
46. Fried O, Tewari A, Zollhöfer M, Finkelstein A, Shechtman E, Goldman DB, Genova K, Jin Z, Theobalt C, Agrawala M (2019) Text-based editing of talking-head video. *ACM Trans Gr* 38(4):1–14
47. Lahiri A, Kwatra V, Frueh C, Lewis J, Bregler C (2021) LipSync3D: data-efficient learning of personalized 3D talking faces from video using pose and lighting normalization. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville
48. Zhang Z, Li L, Ding Y, Fan C (2021) Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR), Nashville
49. Jamaludin A, Chung JS, Zisserman A (2019) You said that?: Synthesizing talking faces from audio. *Int J Comput Vis* 127:1767–1779
50. Choi Y, Choi M, Kim M, Ha J-W, Kim S, Choo J (2018) StarGAN: unified generative adversarial networks for multi-domain image-to-image translation. In: IEEE/CVF conference on computer vision and pattern recognition, Salt Lake City
51. Pumarola A, Agudo A, Martinez AM, Sanfeliu A, Moreno-Noguer F (2019) GANimation: one-shot anatomically consistent facial animation. *Int J Comput Vis* 128:698–713
52. Liu M, Ding Y, Xia M, Liu X, Ding E, Zuo W, Wen S (2019) STGAN: a unified selective transfer network for arbitrary image attribute editing. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach
53. Liang H, Hou X, Shen L (2021) SSFlow: style-guided neural spline flows for face image manipulation. In: Proceedings of the 29th ACM international conference on multimedia, New York
54. Wang R, Chen J, Yu G, Sun L, Yu C, Gao C, Sang N (2021) Attribute-specific Control Units in StyleGAN for Fine-grained image manipulation. In: Proceedings of the 29th ACM international conference on multimedia, New York
55. Karras T, Laine S, Aila T (2019) A style-based generator architecture for generative adversarial networks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach
56. Zhou H, Liu Y, Liu Z, Luo P, Wang X (2019) Talking face generation by adversarially disentangled audio-visual representation. In: AAAI Conference on Artificial Intelligence (AAAI-19), Honolulu

57. Chen L, Maddox RK, Duan Z, Xu C (2019) Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach
58. Vougioukas K, Petridis S, Pantic M (2019) Realistic speech-driven facial animation with GANs. *Int J Comput Vis* 128:1398–1413
59. Thies J, Elgharib M, Tewari A, Theobalt C, Nießner M (2020) Neural voice puppetry: audio-driven facial reenactment. In: European conference on computer vision (ECCV), Glasgow
60. Vougioukas K, Petridis S, Pantic M (2019) End-to-end speech-driven realistic facial animation with temporal GANs. In: Computer Vision and Pattern Recognition (CVPR), Long Beach
61. He Z, Zuo W, Kan M, Shan S, Chen X (2019) AttGAN: facial attribute editing by only changing what you want. *IEEE Trans Image Process* 28(11):5464–5478
62. Shen Y, Gu J, Tang X, Zhou B (2020) Interpreting the latent space of GANs for semantic face editing. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle
63. Jo Y, Park J (2019) SC-FEGAN: face editing generative adversarial network with user's sketch and color. In: IEEE/CVF international conference on computer vision (ICCV), Seoul
64. Shen Y, Yang C, Tang X, Zhou B (2020) InterFaceGAN: interpreting the disentangled face representation learned by GANs. *IEEE Trans Pattern Anal Mach Intell* (Early Access), p 1
65. Fu C, Hu Y, Wu X, Wang G, Zhang Q, He R (2021) High-fidelity face manipulation with extreme poses and expressions. *IEEE Trans Inf Forensics Secur* 16:2218–2231
66. Yang N, Zheng Z, Zhou M, Guo X, Qi L, Wang T (2021) A domain-guided noise-optimization-based inversion method for facial image manipulation. *IEEE Trans Image Process* 30:6198–6211
67. Karras T, Aila T, Laine S, Lehtinen J (2018) Progressive Growing of GANs for improved quality, stability, and variation. In: International conference on learning representations (ICLR), Vancouver
68. Karras T, Laine S, Aittala M, Hellsten J, Lehtinen J, Aila T (2020) Analyzing and improving the image quality of StyleGAN. In: IEEE/CVF Conference on computer vision and pattern recognition (CVPR), Seattle
69. Zhang H, Goodfellow I, Metaxas D, Odena A (2019) Self-attention generative adversarial networks. <http://arxiv.org/abs/1805.08318v2>
70. Brock A, Donahue J, Simonyan K (2019) Self-attention generative adversarial networks. In: International Conference on Learning Representations (ICLR), New Orleans
71. Martin K, Marketing V (2021) What is voice cloning?, ID R&D, <https://www.idrnd.ai/what-is-voice-cloning/>. Accessed 24 July 2021
72. Maheshwari H (2021) Basic text to speech, explained," towards data Science, <https://towardsdatascience.com/text-to-speech-explained-from-basic-498119aa38b5>. Accessed 11 July 2021
73. Maheshwari H (2021) Text to speech system for multi-speaker setting, towards data science, <https://towardsdatascience.com/text-to-speech-system-for-multi-speaker-setting-35e83f84e669>. Accessed 12 July 2021
74. Singh J (2018) WaveNet: google Assistant's voice synthesizer, towardsdatascience, 7 November 2018. <https://towardsdatascience.com/wavenet-google-assistants-voice-synthesizer-a168e9af13b1>. Accessed 10 July 2021
75. Oord AVD, Dieleman S, Zen H, Simonyan K, Vinyals O, Graves A, Kalchbrenner N, Senior A, Kavukcuoglu K (2016) WaveNet: a generative model for raw audio. In: Proceedings of the 9th ISCA Speech Synthesis Workshop, Sunnyvale
76. Oord A, Li Y, Babuschkin I, Simonyan K, Vinyals O, Kavukcuoglu K, Driessche G, Lockhart E, Cobo L, Stimberg F, Casagrande N, Grewe D, Noury S, Dieleman S, Elsen E, Kalchbrenner N, Zen H, Graves A, King H, Walters T, Belov D, Hassabis D (2018) Parallel WaveNet: fast high-fidelity speech synthesis. In: Proceedings of the 35th international conference on machine learning, Stockholm
77. Arik SO, Chrzanowski M, Coates A, Diamos G, Gibiansky A, Kang Y, Li X, Miller J, Ng A, Raiman J, Sengupta S, Shoenybi M (2017) Deep voice: real-time neural text-to-speech. In: International conference on machine learning, Sydney
78. Arik SÖ, Diamos G, Gibiansky A, Miller J, Peng K, Ping W, Raiman J, Zhou Y (2017) Deep voice 2: multi-speaker neural text-to-speech. In: Advances in neural information processing systems, Long Beach
79. Ping W, Peng K, Gibiansky A, Arik SO, Kannan A, Narang S, Raiman J, Miller J (2018) Deep voice 3: scaling text-to-speech with convolutional sequence learning. In: International conference on learning representations (ICLR), Vancouver
80. Wang Y, Skerry-Ryan R, Stanton D, Wu Y, Weiss RJ, Jaitly N, Yang Z, Xiao Y, Chen Z, Bengio S, Le Q, Agiomyrgiannakis Y, Clark R, Saurous RA (2017) Tacotron: towards end-to-end Speech Synthesis. <http://arxiv.org/abs/1703.10135v2>
81. Zhang J-X, Ling Z-H, Liu L-J, Jiang Y, Dai L-R (2019) Sequence-to-sequence acoustic modeling for voice conversion. *IEEE/ACM Trans Audio Speech Lang Process* 27(3):631–644
82. Veaux C, Yamagishi J, King S (2013) Towards personalized synthesized voices for individuals with vocal disabilities: voice banking and reconstruction. In: Speech and language processing for assistive technologies (SLPAT), Grenoble
83. Sisman B, Yamagishi J, King S, Li H (2021) An overview of voice conversion and its challenges: from statistical modeling to deep learning. *IEEE/ACM Trans Audio Speech Lang Process* 29:132–157
84. Zhang J-X, Ling Z-H, Dai L-R (2019) Non-parallel sequence-to-sequence voice conversion with disentangled linguistic and speaker representations. *IEEE/ACM Trans Audio Speech Lang Process* 28:540–552
85. Wang R, Ding Y, Li L, Fan C (2020) One-shot voice conversion using Star-GAN. In: ICASSP 2020 - 2020 IEEE international conference on acoustics, speech and signal processing (ICASSP), Barcelona
86. Liu R, Chen X, Wen X (2020) Voice conversion with transformer network. In: ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP), Barcelona
87. Yasuda Y, Wang X, Takaki S, Yamagishi J (2019) Investigation of enhanced tacotron text-to-speech synthesis systems with self-attention for pitch accent language. In: IEEE international conference on acoustics, speech and signal processing (ICASSP), Brighton
88. Chen Y, Assael Y, Shillingford B, Budden D, Reed S, Zen H, Wang Q, Cobo LC, Trask A, Laurie B, Gulcehre C, Oord AVD, Vinyals O, Freitas ND (2019) Sample efficient adaptive text-to-speech. In: International Conference on Learning Representations (ICLR), New Orleans
89. Liu R, Yang J, Liu M (2019) A new end-to-end long-time speech synthesis system based on Tacotron2. In: International conference proceeding series (ICPS), Beijing
90. Weiss RJ, Skerry-Ryan R, Battenberg E, Mariooryad S, Kingma DP (2021) Wave-Tacotron: spectrogram-free end-to-end text-to-speech synthesis. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto
91. He Q, Xiu Z, Koehler T, Wu J (2021) Multi-rate attention architecture for fast streamable text-to-speech spectrum modeling. In: 2021 IEEE international conference on acoustics, speech and signal processing (ICASSP), Toronto

92. Liu R, Sisman B, Gao G, Li H (2021) Expressive TTS training with frame and style reconstruction loss. *IEEE/ACM Trans Audio Speech Lang Process* 29:1806–1818
93. Zhou X, Ling Z-H, Dai L-R (2021) UnitNet: a sequence-to-sequence acoustic model for concatenative speech synthesis. *IEEE/ACM Trans Audio Speech Lang Process* 29:2643–2655
94. Tanaka K, Kameoka H, Kaneko T, Hojo N (2019) ATTS2S-VC: sequence-to-sequence voice conversion with attention and context preservation mechanisms. In: *ICASSP 2019 - 2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, Brighton
95. Kameoka H, Kaneko T, Tanaka K, Hojo N (2019) ACVAE-VC: non-parallel voice conversion with auxiliary classifier variational autoencoder. *IEEE/ACM Trans Audio Speech Lang Process* 27(9):1432–1443
96. Cong J, Yang S, Xie L, Yu G, Wan G (2020) Data efficient voice cloning from noisy samples with domain adversarial training. In: *Interspeech 2020*, Shanghai
97. Zhang M, Sisman B, Zhao L, Li H (2020) DeepConversion: voice conversion with limited parallel training data. *Speech Commun* 122:31–43
98. Kameoka H, Tanaka K, Kwaśny D, Kaneko T, Hojo N (2020) ConvS2S-VC: fully convolutional sequence-to-sequence voice conversion. *IEEE/ACM Trans Audio Speech Lang Process* 28:1849–1863
99. Ding S, Zhao G, Gutierrez-Osuna R (2020) Improving the speaker identity of non-parallel many-to-many voice conversion with adversarial speaker recognition. In: *INTERSPEECH*, Shanghai
100. Lee S, Ko B, Lee K, Yoo I-C, Yook D (2020) Many-to-many voice conversion using conditional cycle-consistent adversarial networks. In: *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona
101. Zhang M, Zhou Y, Zhao L, Li H (2021) Transfer learning from speech synthesis to voice conversion with non-parallel training data. *IEEE/ACM Trans Audio Speech Lang Process* 29:1290–1302
102. Chen M, Shi Y, Hain T (2021) Towards low-resource stargan voice conversion using weight adaptive instance normalization. In: *IEEE international conference on acoustics, speech and signal processing (ICASSP)*, Toronto
103. Li Z, Tang B, Yin X, Wan Y, Xu L, Shen C, Ma Z (2021) PPG-based singing voice conversion with adversarial representation learning. In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto
104. Kameoka H, Huang W-C, Tanaka K, Kaneko T, Hojo N, Toda T (2021) Many-to-many voice transformer network. *IEEE/ACM Trans Audio Speech Lang Process* 29:656–670
105. Li H, Li B, Tana S, Huang J (2020) Identification of deep network generated images using disparities in color components. *Signal Process* 174:107616
106. Chen P, Liu J, Liang T, Yu C, Zou S, Dai J, Han J (2021) DLFM-Net: end-to-end detection and localization of face manipulation using multi-domain features. In: *IEEE international conference on multimedia and expo (ICME)*, Shenzhen
107. McCloskey S, Albright M (2018) Detecting GAN-generated imagery using color cues. <http://arxiv.org/abs/1812.08247v1>
108. Yu N, Davis L, Fritz M (2019) Attributing fake images to GANs: learning and analyzing GAN fingerprints. In: *IEEE/CVF international conference on computer vision (ICCV)*, Seoul
109. Koopman M, Rodriguez AM, Geradts Z (2018) Detection of deepfake video manipulation. In: *Irish machine vision and image processing conference (IMVIP)*, Belfast
110. Li Y, Lyu S (2019) Exposing DeepFake videos by detecting face warping artifacts. In: *IEEE/CVF conference on computer vision and pattern recognition (CVPR) workshops*, Long Beach
111. Li L, Bao J, Zhang T, Yang H, Chen D, Wen F, Guo B (2020) Face X-ray for more general face forgery detection. In: *IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, Seattle
112. Matern F, Riess C, Stamminger M (2019) Exploiting visual artifacts to expose deepfakes and face manipulations. In: *IEEE winter applications of computer vision workshops (WACVW)*, Waikoloa
113. Zhao Y, Ge W, Li W, Wang R, Zhao L, Ming J (2019) Capturing the persistence of facial expression features for deepfake video detection. In: *International Conference on Information and Communications Security*, Beijing
114. Li X, Yu K, Ji S, Wang Y, Wu C, Xue H (2020) Fighting against deepfake: Patch&Pair convolutional neural networks (PPCNN). In: *Companion Proceedings of the Web Conference 2020*, New York
115. Lee S, Tariq S, Shin Y, Woo SS (2021) Detecting handcrafted facial image manipulations and GAN-generated facial images using Shallow-FakeFaceNet. *Appl Soft Comput* 105:107256
116. Shang Z, Xie H, Zha Z, Yu L, Li Y, Zhang Y (2021) PRRNet: Pixel-Region relation network for face forgery detection. *Pattern Recognit* 116:107950
117. Agarwal S, Farid H, Fried O, Agrawala M (2020) Detecting deepfake videos from phoneme-viseme mismatches. In: *IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW)*, Seattle
118. Mittal T, Bhattacharya U, Chandra R, Bera A, Manocha D (2020) Emotions don't lie: an audio-visual deepfake detection method using affective cues. In: *ACM international conference on multimedia*, New York
119. Chugh K, Gupta P, Dhali A, Subramanian R (2020) Not made for each other- audio-visual dissonance-based deepfake detection and localization. In: *ACM international conference on multimedia*, New York
120. Hosier BC, Stamm MC (2020) Detecting video speed manipulation. In: *IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW)*, Seattle
121. Amerini I, Galteri L, Caldelli R, Bimbo AD (2019) Deepfake video detection through optical flow based CNN. In: *IEEE/CVF international conference on computer vision workshop (ICCVW)*, Seoul.
122. Caldelli R, Galteri L, Amerini I, Bimbo AD (2021) Optical Flow based CNN for detection of unlearned deepfake manipulations. *Pattern Recognit Lett* 146:31–37
123. Yang X, Li Y, Lyu S (2019) Exposing deep fakes using inconsistent head poses. In: *IEEE international conference on acoustics, speech and signal processing (ICASSP)*, Brighton
124. Li Y, Chang M-C, Lyu S (2018) In Ictu Oculi: exposing AI created fake videos by detecting eye blinking. In: *IEEE International Workshop on Information Forensics and Security (WIFS)*, Hong Kong
125. Qi H, Guo Q, Juefei-Xu F, Xie2 X, Ma L, Feng W, Liu Y, Zhao J (2020) DeepRhythm: exposing DeepFakes with attentional visual heartbeat rhythms. In: *ACM international conference on multimedia*, New York
126. Ciftci UA, Demir I, Yin L (2020) FakeCatcher: detection of synthetic portrait videos using biological signals. *IEEE Trans Pattern Anal Mach Intell* (Early Access)
127. Hernandez-Ortega J, Tolosana R, Fierrez J, Morales A (2020) DeepFakesON-Phys: deepfakes detection based on heart rate estimation. <http://arxiv.org/abs/2010.00400v3>
128. Yasrab R, Jiang W, Riaz A (2021) Fighting deepfakes using body language analysis. *Forecast MDPI Open Access J* 3(2):1–19
129. Khalid H, Woo SS (2020) OC-FakeDect: classifying deepfakes using one-class variational Autoencoder. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, Seattle

130. Xuan X, Peng B, Wang W, Dong J (2019) On the generalization of GAN image forensics. In: Chinese conference on biometric recognition, Zhuzhou
131. Zhou P, Han X, Morariu VI, Davis LS (2017) Two-stream neural networks for tampered face detection. In: IEEE conference on computer vision and pattern recognition workshops (CVPRW), Honolulu
132. Jeon H, Bang Y, Woo SS (2019) FakeTalkerDetect: effective and practical realistic neural talking head detection with a highly unbalanced dataset. In: IEEE/CVF international conference on computer vision workshop (ICCVW), Seoul
133. Wu X, Xie Z, Gao Y, Xiao Y (2020) SSTNet: detecting manipulated faces through spatial, steganalysis and temporal features. In: IEEE International conference on acoustics, speech and signal processing (ICASSP), Barcelona
134. Tariq S, Lee S, Kim H, Shin Y, Woo SS (2019) GAN is a friend or foe? A framework to detect various fake face images. In: Proceedings of the 34th ACM/SIGAPP symposium on applied computing, Cyprus
135. Sohrawardi SJ, Chintha A, Thai B, Seng S, Hickerson A, Ptucha R, Wright MK (2019) Poster: towards robust open-world detection of deepfakes. In: ACM SIGSAC conference on computer and communications security, London
136. Fernando T, Fookes C, Denman S, Sridharan S (2019) Exploiting human social cognition for the detection of fake and fraudulent faces via memory networks. <http://arxiv.org/abs/1911.07844v1>
137. Sun X, Wu B, Chen W (2020) Identifying invariant texture violation for robust deepfake detection. <http://arxiv.org/abs/2012.10580v1>
138. Ding X, Raziei Z, Larson EC, Olinick EV, Krueger P, Hahsler M (2020) Swapped face detection using deep learning and subjective assessment. EURASIP J Inf Secur, vol. 6
139. Kumar A, Bhavsar A, Verma R (2020) Detecting deepfakes with metric learning. In: International Workshop on Biometrics and Forensics (IWBF), Porto
140. Rana MS, Sung AH (2020) DeepfakeStack: a deep ensemble-based learning technique for deepfake detection. In: IEEE international conference on cyber security and cloud computing, New York
141. Zhou X, Wang Y, Wu P (2020) Detecting deepfake videos via frame serialization learning. In: IEEE 3rd International Conference of Safe Production and Informatization (IICSPI), Chongqing City
142. Nguyen XH, Tran TS, Le VT, Nguyen KD, Truong D-T (2021) Learning Spatio-temporal features to detect manipulated facial videos created by the Deepfake techniques. Forensic Sci Int Digital Investig 36:301108
143. Xu Z, Liu J, Lu W, Xu B, Zhao X, Li B, Huang J (2021) Detecting facial manipulated videos based on set convolutional neural networks. J Vis Commun Image Represent 77:103119
144. Chen Z, Yang H (2021) Attentive semantic exploring for manipulated face detection. In: IEEE International Conference on Acoustics, Speech and Signal Processing, Toronto
145. Zhang J, Ni J, Xie H (2021) DeepFake videos detection using self-supervised decoupling network. In: IEEE International Conference on Multimedia and Expo (ICME), Shenzhen
146. Gu Z, Chen Y, Yao T, Ding S, Li J, Huang F, Ma L (2021) Spatiotemporal inconsistency learning for deepfake video detection. In: Proceedings of the 29th ACM international conference on multimedia, New York
147. Tu Y, Liu Y, Li X (2021) Deepfake video detection by using convolutional gated recurrent unit. In: International conference on machine learning and computing, Shenzhen
148. Zhuang Y-X, Hsu C-C (2019) Detecting generated image based on a coupled network with two-step pairwise learning. In: IEEE international conference on image processing (ICIP), Taipei
149. Lima OD, Franklin S, Basu S, Karwoski B, George A (2020) Deepfake detection using spatiotemporal convolutional networks. <http://arxiv.org/abs/2006.14749v1>
150. Lang Y, Li X, Chen Y, Mao X, He Y, Wang S, Xue H, Lu Q (2020) Sharp multiple instance learning for deepfake video detection. In: Proceedings of the 28th ACM international conference on multimedia, Seattle WA
151. Chen B, Ju X, Xiao B, Ding W, Zheng Y, Albuquerque VHCD (2021) Locally GAN-generated face detection based on an improved Xception. Inf Sci 572:16–28
152. Chen H-S, Rouhsedaghat M, Ghani H, Hu S, You S, Kuo C-CJ (2021) DefakeHop: a light-weight high-performance deepfake detector. In: IEEE International Conference on Multimedia and Expo (ICME), Shenzhen
153. Das S, Seferbekov S, Datta A, Islam MS, Amin MR (2021) Towards solving the deepfake problem : an analysis on improving deepfake detection using dynamic face augmentation. In: IEEE/CVF international conference on computer vision workshops (ICCVW), Montreal
154. Nguyen HH, Fang F, Yamagishi J, Echizen I (2019) Multi-task learning for detecting and segmenting manipulated facial images and videos. In: IEEE 10th international conference on biometrics theory, applications and systems (BTAS), Tampa
155. Du M, Pentyala SK, Li Y, Hu X (2020) Towards generalizable deepfake detection with locality-aware autoencoder. In: ACM international conference on information & knowledge management, Virtual Event Ireland
156. He P, Li H, Wang H (2019) Detection of fake images via the ensemble of deep representations from multi color spaces. In: IEEE International conference on image processing (ICIP), Taipei
157. Guo Z, Yang G, Chen J, Sun X (2021) Fake face detection via adaptive manipulation traces extraction network. Comput Vis Image Underst 204:103170
158. Wang R, Juefei-Xu F, Ma L, Xie X, Huang Y, Wang J, Liu Y (2020) FakeSpotter: a simple yet robust baseline for spotting AI-synthesized fake faces. In: International joint conference on artificial intelligence (IJCAI), Yokohama
159. Khan SA, Dai H (2021) Video transformer for deepfake detection with incremental learning. In: Proceedings of the 29th ACM international conference on multimedia, New York
160. Frank J, Eisenhofer T, Schonherr L, Fischer A, Kolossa D, Holz T (2020) Leveraging frequency analysis for deep fake image recognition. Proc of Mach Learn 119:3247–3258
161. Durall R, Keuper M, Pfrendt F-J, Keuper J (2020) Unmasking deepfakes with simple feature. <http://arxiv.org/abs/1911.00686v3>
162. Masi I, Killekar A, Mascarenha RM, Gurudatt SP, AbdAlmageed W (2020) Two-branch recurrent network for isolating deepfakes in videos. In: European conference on computer vision, Glasgow
163. McCloskey S, Albright M (2019) Detecting GAN-generated imagery using saturation cues. In: IEEE International conference on image processing (ICIP), Taipei
164. Guarnera L, Giudice O, Battato S (2020) DeepFake detection by analyzing convolutional traces. In: IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW), Seattle
165. Wang S-Y, Wang O, Zhang R, Owens A, Efros AA (2020) CNN-generated images are surprisingly easy to spot... for now. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR), Seattle
166. Lugstein F, Baier S, Bachinger G, Uhl A (2021) PRNU-based deepfake detection. In: Proceedings of the 2021 ACM workshop on information hiding and multimedia security
167. Nirkin Y, Wolf L, Keller Y, Hassner T (2020) DeepFake detection based on discrepancies between faces and their context. <http://arxiv.org/abs/2008.12262v1>

168. Yang J, Xiao S, Li A, Lan G, Wang H (2021) Detecting fake images by identifying potential texture difference. *Futur Gener Comput Syst* 125:127–135
169. Li G, Cao Y, Zhao X (2021) Exploiting facial symmetry to expose deepfakes. In: *IEEE international conference on image processing (ICIP)*, Anchorage
170. Luo Z, Kamata S-I, Sun Z (2021) Transformer and node-compressed dnn based dual-path system for manipulated face detection. In: *IEEE international conference on image processing (ICIP)*, Anchorage
171. Yang J, Xiao S, Li A, Lu W, Gao X, Li Y (2021) MSTA-net: forgery detection by generating manipulation trace based on multi-scale self-texture attention. *IEEE Trans Circuits Syst Video Technol* (Early Access), pp. 1–1
172. Bonomi M, Pasquini C, Boato G (2021) Dynamic texture analysis for detecting fake faces in video sequences. *J Vis Commun Image Represent* 79:103239
173. Yang J, Li A, Xiao S, Lu W, Gao X (2021) MTD-Net: learning to detect deepfakes images by multi-scale texture difference. *IEEE Trans Inf Forensics Secur* 16:4234–4245
174. Gu Y, He M, Nagano K, Li H (2019) Protecting world leaders against deep fakes. In: *IEEE/CVF conference on computer vision and pattern recognition (CVPR) Workshops*, Long Beach
175. Yang C-Z, Ma J, Wang S-L, Liew AW-C (2020) Preventing deepfake attacks on speaker authentication by dynamic lip movement analysis. *IEEE Trans Inf Forensics Secur* 16:1841–1854
176. Hosler B, Salvi D, Murray A, Antonacci F, Bestagini P, Tubaro S, Stamm MC (2021) Do deepfakes feel emotions? A semantic approach to detecting deepfakes via emotional inconsistencies. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Nashville
177. Demir İ, Ciftci UA (2021) Where do deep fakes look? Synthetic face detection via gaze. In *ACM symposium on eye tracking research and applications*, Germany
178. Hu S, Li Y, Lyu S (2021) Exposing GAN-generated faces using inconsistent corneal specular highlights. In: *IEEE international conference on acoustics, speech and signal processing (ICASSP)*, Toronto
179. Agarwal S, Farid H (2021) Detecting deep-fake videos from aural and oral dynamics. In: *IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW)*, Nashville
180. Sabir E, Cheng J, Jaiswal A, AbdAlmageed W, Masi I, Natarajan P (2019) Recurrent-convolution approach to deepfake detection – state-of-art results on FaceForensics++. <http://arxiv.org/abs/1905.00582v1>
181. Sabir E, Cheng J, Jaiswal A, AbdAlmageed W, Masi I, Natarajan P (2019) Recurrent convolutional strategies for face manipulation detection in videos. In: *IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, Long Beach
182. Amerini I, Caldelli R (2020) Exploiting prediction error inconsistencies through LSTM-based classifiers to detect deepfake videos. In: *ACM workshop on information hiding and multimedia security*, New York
183. Lu C, Liu B, Zhou W, Chu Q, Yu N (2021) Deepfake video detection using 3D-attentional inception convolutional neural network. In: *IEEE international conference on image processing (ICIP)*, Anchorage
184. Trinh L, Tsang M, Rambhatla S, Liu Y (2021) Interpretable and trustworthy deepfake detection via dynamic prototypes. In: *IEEE winter conference on applications of computer vision (WACV)*, Hawaii
185. Cozzolino D, Thies J, Rossler A, Riess C, Nießner M, Verdoliva L (2019) ForensicTransfer: weakly-supervised domain adaptation for forgery detection. <http://arxiv.org/abs/1812.s02510v2>
186. Hsu C-C, Zhuang Y-X, Lee C-Y (2019) Deep fake image detection based on pairwise learning. *Appl Sci* 10(1):370
187. Dang LM, Hassan SI, Im S, Moon H (2019) Face image manipulation detection based on a convolutional neural network. *Expert Syst Appl* 129:156–168
188. Nguyen HH, Yamagishi J, Echizen I (2019) Capsule-forensics: using capsule networks to detect forged images and videos. In: *IEEE international conference on acoustics, speech and signal processing (ICASSP)*, Brighton
189. Montserrat DM, Hao H, Yarlagadda SK, Baireddy S, Shao R, Horváth J, Bartusiak E, Yang J, Güera D, Zhu F, Delp EJ (2020) Deepfakes detection with automatic face weighting. In: *IEEE/CVF conference on computer vision and pattern recognition workshops*, Seattle
190. Choi DH, Lee HJ, Lee S, Kim JU, Ro YM (2020) Fake video detection with certainty-based attention network. In: *IEEE international conference on image processing (ICIP)*, Abu Dhabi
191. Chintha A, Thai B, Sohrwardi SJ, Bhatt K, Hickerson A, Wright M, Ptucha R (2020) Recurrent convolutional structures for audio spoof and video deepfake detection. *IEEE J Sel Top Signal Process* 14(5):1024–1037
192. Hu J, Wang S, Li X (2021) Improving the generalization ability of deepfake detection via disentangled representation learning. In: *IEEE international conference on image processing (ICIP)*, Anchorage
193. Hu J, Liao X, Wang W, Qin Z (2021) Detecting compressed deepfake videos in social networks using frame-temporality two-stream convolutional network. *IEEE Trans Circuits Syst Video Technol* (Early Acces) 32(3):1089–1102
194. Han B, Han X, Zhang H, Li J, Cao X (2021) Fighting fake news: two stream network for deepfake detection via learnable SRM. *IEEE Trans Biometrics Behav Ident Sci* 3(3):320–331
195. Kim M, Tariq S, Woo SS (2021) FReTAL: generalizing deepfake detection using knowledge distillation and representation learning. In: *IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW)*, Nashville
196. Zhao H, Wei T, Zhou W, Zhang W, Chen D, Yu N (2021) Multi-attentional deepfake detection. In: *IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, Nashville
197. Sun Z, Han Y, Hua Z, Ruan N, Jia W (2021) Improving the efficiency and robustness of deepfakes detection through precise geometric features. In: *IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, Nashville
198. Tariq S, Lee S, Woo SS (2021) One detector to rule them all. In: *Proceedings of the web conference 2021*, New York
199. Wang R, Juefei-Xu F, Huang Y, Guo Q, Xie X, Ma L, Liu Y (2020) DeepSonar: towards effective and robust detection of AI-synthesized fake voices. In: *Proceedings of the 28th ACM international conference on multimedia*, Seattle
200. Balamurli B, Lin KE, Lui S, Chen J-M, Herremans D (2019) Toward robust audio spoofing detection: a detailed comparison of traditional and learned features. In: *IEEE Access*
201. Saranya MS, Padmanabhan R, Murthy HA (2018) Replay attack detection in speaker verification using non-voiced segments and decision level feature switching. In: *International conference on signal processing and communications (SPCOM)*, Bangalore
202. Witkowski M, Kacprzak S, Zelasko P, Kowalczyk K, Gałka J (2017) Audio replay attack detection using high-frequency features. In: *INTERSPEECH*, Stockholm
203. AlBadawy EA, Lyu S, Farid H (2019) Detecting AI-synthesized speech using bispectral analysis. In: *IEEE/CVF Conference on computer vision and pattern recognition (CVPR)*, Long Beach
204. Patil HA, Kamble MR (2018) A survey on replay attack detection for automatic speaker verification (ASV) system. In: *Proceedings of the APSIPA Annual Summit and Conference 2018*, Hawaii
205. Wijethunga R, Matheesha D, Noman AA, Silva KD, Tissera M, Rupasinghe L (2020) Deepfake audio detection: a deep learning

- based solution for group conversations. In: International conference on advancements in computing (ICAC), Malabe
206. Chen T, Kumar A, Nagarsheth P, Sivaraman G, Khoury E (2020) Generalization of audio deepfake detection. In: Odyssey 2020 the speaker and language recognition workshop, Tokyo
 207. Shim H-J, Jung J-W, Heo H-S, Yoon S-H, Yu H-J (2018) Replay spoofing detection system for automatic speaker verification using multi-task learning of noise classes. In: Conference on technologies and applications of artificial intelligence (TAAI), Taichung
 208. Yang J, Das RK (2020) Long-term high frequency features for synthetic speech detection. *Digital Signal Process* 97:102622
 209. Malik H (2019) Securing voice-driven interfaces against Fake (Cloned) Audio Attacks. In: IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), San Jose
 210. Gunendradasan T, Irtza S, Ambikairajah E, Epps J (2019) Transmission line cochlear model based AM-FM features for replay attack detection. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton
 211. Borrelli C, Bestagini P, Antonacci F, Sarti A, Tubaro S (2021) Synthetic speech detection through short-term and long-term prediction traces. *EURASIP J Inf Secur*, 2
 212. Lai C-I, Abad A, Richmond K, Yamagishi J, Dehak N, King S (2019) Attentive filtering networks for audio replay attack detection. In: IEEE international conference on acoustics, speech and signal processing (ICASSP), Brighton
 213. Huang L, Pun C-M (2019) Audio replay spoof attack detection using segment-based hybrid feature and DenseNet-LSTM network. In: IEEE international conference on acoustics, speech and signal processing (ICASSP), Brighton
 214. Gomez-Alanis A, Peinado AM, Gonzalez JA, Gomez AM (2019) A light convolutional GRU-RNN deep feature extractor for ASV spoofing detection. In: INTERSPEECH, Graz
 215. Gomez-Alanis A, Peinado AM, Gonzalez JA, Gomez AM (2021) A gated recurrent convolutional neural network for robust spoofing detection. *IEEE/ACM Trans Audio Speech Lang Process* 27(12):1985–1999
 216. Huang L, Pun C-M (2020) Audio replay spoof attack detection by joint segment-based linear filter bank feature extraction and attention-enhanced DenseNet-BiLSTM Network. *IEEE/ACM Trans Audio Speech Lang Process* 28:1813–1825
 217. Wu Z, Das RK, Yang J, Li H (2020) Light convolutional neural network with feature genuinization for detection of synthetic speech attacks. In: INTERSPEECH, Shanghai
 218. Wang Z, Cui S, Kang X, Sun W, Li Z (2021) Densely connected convolutional network for audio spoofing detection. In: Asia-Pacific signal and information processing association annual summit and conference (APSIPA ASC), Auckland
 219. You CH, Yang J (2020) Device feature extraction based on parallel neural network training for replay spoofing detection. *IEEE/ACM Trans Audio Speech Lang Process* 28:2308–2318
 220. Luo A, Li E, Liu Y, Kang X, Wang ZJ (2021) A capsule network based approach for detection of audio spoofing attacks. In: IEEE International conference on acoustics, speech and signal processing (ICASSP), Toronto
 221. Ren Y, Liu W, Liu D, Wang L (2021) Recalibrated bandpass filtering on temporal waveform for audio spoof detection. In: IEEE International conference on image processing (ICIP), Anchorage
 222. Huang L, Zhao J (2021) Audio replay spoofing attack detection using deep learning feature and long-short-term memory recurrent neural network. In: The second international conference on artificial intelligence, information processing and cloud computing, Hangzhou
 223. Ouyang M, Das RK, Yang J, Li H (2021) Capsule network based end-to-end system for detection of replay attacks. In: International symposium on chinese spoken language processing (ISCSLP), Hong Kong
 224. Li X, Li N, Weng C, Liu X, Su D, Yu D, Meng H (2021) Replay and synthetic speech detection with Res2Net architecture. In: IEEE International conference on acoustics, speech and signal processing (ICASSP), Toronto
 225. Li Y, Yang X, Sun P, Qi H, Lyu S (2020) Celeb-DF: a large-scale challenging dataset for deepfake forensics. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle
 226. Dolhansky B, Bitton J, Pflaum B, Lu J, Howes R, Wang M, Ferrer C C (2020) The deepfake detection challenge (DFDC) dataset. In: <http://arxiv.org/2006.07397v4>
 227. Korshunov P, Marcel S (2018) DeepFakes: a new threat to face recognition? Assessment and Detection. In: <http://arxiv.org/1812.08685v1>
 228. Rossler A, Cozzolino D, Verdoliva L, Riess C, Thies J, Nießner M (2018) FaceForensics: a large-scale video dataset for forgery detection in human faces. <http://arxiv.org/1803.09179v1>
 229. Khodabakhsh A, Ramachandra R, Raja K, Wasnik P, Busch C (2018) Fake face detection methods: can they be generalized? In: International Conference of the Biometrics Special Interest Group (BIOSIG), Darmstadt
 230. Dolhansky B, Howes R, Pflaum B, Baram N, Ferrer CC (2019) The deepfake detection challenge (DFDC) preview dataset. In: <http://arxiv.org/1910.08854v2>
 231. Contributing Data to Deepfake Detection Research, (2019). <https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html>
 232. Jiang L, Li R, Wu W, Qian C, Loy CC (2020) DeeperForensics-1.0: a large-scale dataset for real-world face forgery detection. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle
 233. Zi B, Chang M, Chen J, Ma X, Jiang Y-G (2020) WildDeepfake: a challenging real-world dataset for deepfake detection. In: Proceedings of the 28th ACM international conference on multimedia, Seattle
 234. Dong X, Bao J, Chen D, Zhang W, Yu N, Chen D, Wen F, Guo B (2020) Identity-driven deepfake detection. In: <http://arxiv.org/2012.03930v1>
 235. Huang J, Wang X, Du B, Du P, Xu C (2021) DeepFake MNIST+: a DeepFake facial animation dataset. In: IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Montreal
 236. Kominek J, Black AW (2004) The CMU Arctic speech databases. In: Fifth ISCA Workshop on Speech Synthesis
 237. Panayotov V, Chen G, Povey D, Khudanpur S (2015) Librispeech: an ASR corpus based on public domain audio books. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane
 238. Wu Z, Kinnunen T, Evans N, Yamagishi J, Hanilc C, Sahidullah IM, Sizov A (2015) ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge. In: InterSpeech, Dresden
 239. Ito K, Johnson L (2017) The LJ speech dataset, LibriVox project. <https://keithito.com/LJ-Speech-Dataset/>. Accessed 28 July 2021
 240. Delgado H, Todisco M, Sahidullah M, Evans N, Kinnunen T, Lee KA, Yamagishi J (2018) ASVspoof 2017 Version 2.0: metadata analysis and baseline enhancements. In: Odyssey 2018—the speaker and language recognition workshop, Les Sables
 241. Chung JS, Nagrani A, Zisserman A (2018) VoxCeleb2: Deep speaker recognition. In: INTERSPEECH, Hyderabad
 242. Veaux C, Yamagishi J, MacDonald K (2019) CSTR VCTK Corpus: English multi-speaker Corpus for CSTR voice cloning toolkit. The Centre for Speech Technology Research (CSTR), University of Edinburgh
 243. Reimao R, Tzerpos V (2019) FoR: a dataset for synthetic speech detection. In: International conference on speech technology and human-computer dialogue (SpeD), Timisoara

244. Nagrani A, Chung JS, Xie W, Zisserman A (2020) Voxceleb: Large-scale speaker verification in the wild. *Comput Speech Lang* 60:101027S
245. GMAIL. The M-AILABS Speech dataset, Caito, <https://www.caito.de/2019/01/the-m-ailabs-speech-dataset/>. Accessed 28 July 2021].
246. Wang X, Yamagishi J, Todisco M, Delgado H, Nautsch A, Evans N, Sahidullah M, Vestman V, Kinnunen T, Lee KA, Juvela L, Alku P, Peng Y-H, Hwang H-T, Tsao Y, Wang H-M, Maguer SL, Becker M, Henderson F, Clark R, Zhang Y, Wang Q, Jia Y, Onuma K, Mushika K, Kaneda T, Jiang Y, Liu L-J, Wu Y-C, Huang W-C, Toda T, Tanaka K, Kameoka H, Steiner I, Matrouf D, Bonastre J-F, Govender A, Ronanki S, Zhang J-X, Ling Z-H (2020) ASVspoof 2019: a large-scale public database of synthesized, converted and replayed speech. *Comput Speech Lang* 64:101114
247. Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. In: *Advances in neural information processing systems*, Montreal
248. DepFaceLab, GitHub, [Online]. Available: <https://github.com/iperov/DeepFaceLab>. Accessed 6 April 2021
249. Deepfakes web, [Online]. Available: <https://deepfakesweb.com/>. Accessed 6 April 2021
250. FaceApp, [Online]. Available: <https://www.faceapp.com/>. Accessed 1 April 2021
251. Zao, [Online]. Available: <https://zaodownload.com/>. Accessed 6 April 2021
252. MachineTube, [Online]. Available: <https://www.machine.tube/>. Accessed 6 April 2021
253. Doublicat, [Online]. Available: <https://reface.app/about/>. Accessed 7 April 2021
254. Resemble AI, [Online]. Available: <https://www.resemble.ai/>. Accessed 28 08 2021
255. Rudrabha/Wav2Lip, github, [Online]. Available: <https://github.com/Rudrabha/Wav2Lip>
256. Thies J, Zollhöfer M, Theobalt C, Stamminger M, Nießner M (2018) Headon: real-time reenactment of human portrait videos. *ACM Trans Gr* 37(4):1–13
257. Nguyen TT, Nguyen CM, Nguyen DT, Nguyen DT, Nahavandi S (2019) Deep learning for deepfakes creation and detection <http://arxiv.org/1909.11573v1>

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