REVIEW ARTICLE

A Comprehensive Survey of DeepFake Generation and Detection Techniques in Audio-Visual Media

Iqra Khan[1], **Kashif Khan**[2] **and Arshad Ahmad**[3,*]

[1] Department of Computer and Software Technology, University of Swat, Swat 19130, Pakistan
[2] Faculty of Computing, Universiti Malaysia Pahang Al-Sultan Abdullah, 26600 Pekan, Malaysia
[3] Department of Computer Software Engineering, Military College of Signals (MCS), National University of Sciences and Technology (NUST), Rawalpindi 46000, Pakistan

Abstract

The rapid advancement in machine learning and artificial intelligence has significantly enhanced capabilities in multimedia content creation, particularly in the domain of deepfake generation. Deepfakes leverage complex neural networks to create hyper-realistic manipulated audio-visual content, raising profound ethical, societal, and security concerns. This paper presents a comprehensive survey of contemporary trends in deepfake video research, focusing on both generation and detection methodologies. The study categorizes deepfakes into three primary types: facial manipulation, lip-synchronization, and audio deepfakes, further subdividing them into face swapping, face generation, attribute manipulation, puppeteering, speech generation, and voice conversion. For each type, the paper reviews cutting-edge generation techniques, including StyleGANs, variational autoencoders, and various speech synthesis models. It also presents an in-depth analysis of detection methods, highlighting both traditional handcrafted feature-based approaches and modern deep learning frameworks utilizing CNNs, RNNs, attention mechanisms, and hybrid transformer models. The paper evaluates these methods in terms of performance, generalization, robustness, and limitations against evolving deepfake techniques. The survey identifies significant challenges such as vulnerability to adversarial attacks, lack of generalized models, and dependency on high-quality training data. The insights provided aim to aid researchers and practitioners in developing more robust detection mechanisms and understanding the landscape of deepfake threats and countermeasures. Ultimately, this study contributes to the growing body of literature by mapping current trends and suggesting potential avenues for future research in combating deepfake proliferation.

Keywords: DeepFake, deep learning, facial manipulations, puppeteering, lip-synchronization, image processing.

Submitted: 07 May 2025
Accepted: 24 June 2025
Published: 30 June 2025

Vol. 1, **No.** 2, 2025.
10.62762/JIAP.2025.431672

***Corresponding author:**
✉ Arshad Ahmad
arshad.ahmad@mcs.nust.edu.pk

Citation
Khan, I., Khan, K., & Ahmad, A. (2025). A Comprehensive Survey of DeepFake Generation and Detection Techniques in Audio-Visual Media. *ICCK Journal of Image Analysis and Processing*, 1(2), 73–95.

1 Introduction

The advent of personal computing has democratized video editing, making it widely accessible to the general public. The fields of photographic, audio, and video content creation and editing have undergone transformative changes due to advancements in artificial intelligence (AI) and machine learning (ML) algorithms. These technologies have had a profound impact on image processing, enabling capabilities such as enhancing image quality, noise reduction, color correction, and the generation of realistic visual effects [16]. The sophistication and accessibility of image processing techniques have significantly evolved, empowering both professionals and amateurs with tools that range from simple filters and modifications to advanced object removal and scene reconstruction. However, the rise of deepfake videos has introduced ethical dilemmas and potential risks. Deepfakes, which can generate highly realistic but fraudulent content, raise serious concerns about misinformation, declining public trust in visual media, and misuse for malicious purposes such as fabricated evidence, revenge pornography, and political manipulation [82].

Numerous applications and software have emerged, enabling the public to create realistic deepfake images and videos. These tools pose substantial risks, including undermining the credibility of information sources, disrupting political processes, and harming individuals. For instance, in 2017, a Reddit user named "Deepfake" initiated a forum for distributing pornographic content created by swapping actors' faces, leading to significant reputational damage for public figures [62]. Similarly, in 2018, BuzzFeed showcased a video of Barack Obama delivering a public service announcement about deepfake technology, but the voice and expressions were generated through actor Jordan Peele's audio manipulation [28]. Since then, deepfake technology has advanced rapidly, becoming increasingly accessible through free smartphone applications such as FakeApp [2], ReFace [7], and ZAO [8]. Dedicated software like FaceSwap [3] and DeepFaceLab [4] further enables enthusiasts to create professional-grade fabricated visuals and audio.

In addition to altering visual content, audio deepfakes have also demonstrated their potential for harm. Advanced speech synthesis techniques, including Tacotron [154], Deep Voice [20], and WaveNet [112], have made it easier to produce highly realistic fake voices. Such technologies have already facilitated significant incidents, such as a 2019 financial fraud where a CEO was tricked into transferring $243,000 by an audio deepfake imitating their voice [52]. The ability to impersonate influential voices poses severe threats to national security [19].

This paper aims to provide a comprehensive comparative analysis of state-of-the-art deepfake generation and detection systems. By examining these systems' performance across various categories, this study seeks to enhance the understanding of deepfake technology and identify potential areas for improvement to mitigate associated risks effectively.

This manuscript provides a comprehensive and structured review of the current landscape in deepfake technology, with key contributions outlined as follows:

- Categorization of Deepfakes: Introduced a structured classification into facial manipulations, lip-synchronization, and audio deepfakes, with further granular distinctions (e.g., face swapping, voice conversion).
- Survey of Generation Techniques: Reviewed state-of-the-art generation models including StyleGAN, WaveNet, Tacotron, and other GAN-based architectures used for both image and audio manipulation.
- Survey of Detection Techniques: Analyzed both traditional (e.g., SVM, optical flow) and deep-learning-based detection techniques (e.g., CNNs, RNNs, transformers, hybrid models).
- Comparative Analysis: Compared performance metrics, dataset usage, robustness, and vulnerabilities of various generation and detection methods.
- Identification of Limitations: Highlighted challenges such as generalization to unseen data, adversarial resistance, and data dependence.
- Future Research Directions: Proposed the need for multi-modal detection, larger and more diverse datasets, adversarially robust models, and real-time detection systems.
- Practical Relevance: Provided use cases and real-world examples (e.g., fake political videos, financial scams) to emphasize the urgency of effective countermeasures.

2 Audio-Visual Deepfakes Categorization

Deepfake videos, a product of advanced machine learning techniques, expose various categorizations

according to their unique characteristics and intended purposes. Various types of deepfake videos, each performing a particular function or producing a specific effect using advanced algorithms to modify and synthesize visual and audio content. The visual representation of data on mobile devices to manipulate information is also important that can further enhanced to interact with visual deepfake [68, 71]. The categorization is presented in Figure 1 and briefly discussed below:

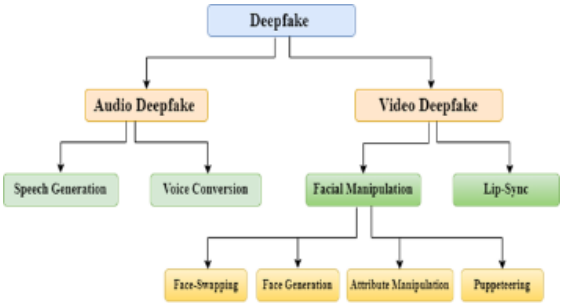

Figure 1. Audio-visual Deepfake categorization.

2.1 Facial Manipulation

Manipulation of facial features, each of which focuses on different aspects of facial transformation, can be grouped into four main categories, such as:

1. **Face Swapping**

   Face Swapping, often called face exchange, is a technique that exchanges one individual's face with another individual in an image or video. This technique analyzes and modifies facial features using advanced algorithms and deep learning models, resulting in a seamless and realistic transformation. Applications like FakeApp [2], ZAO [8], ReFace [7], and FaceSwap [3] have become well-known for producing convincing Face swapping-based deepfake.

2. **Face Generation**

   Face generation refers to generating realistic and aesthetic human faces that cannot be found in reality but have natural and human-like features. This technique generates entire face images using VAEs [80] and GANs [44] such as the StyleGAN [65] method. VAE, also known as variational autoencoders, can learn the underlying latent space of facial features from a dataset.

3. **Attributes Manipulation**

   Face attribute manipulation involves changing certain facial features, such as modifying the person's age, hair, gender, or skin color, and the inclusion of eyewear. This method frequently makes use of GAN, such as StarGAN [34]. FaceApp [1] and other publicly available AI-based face editing software have grown in popularity because they provide user-friendly tools for changing the appearance of facial features in images.

4. **Puppeteering**

   Puppeteering referred to as face reenactment, is a method during which the movement and expression of an individual in a source video are passed to the target video, giving the impression that the individual in the target video is being controlled like a puppet by the action of the individual in the source video. However, this category's most frequently used technique is Neural texture [141] and Face2Face [142].

2.2 lip-synchronization

The technology of lip-synchronization deepfakes is a remarkable development in artificial intelligence and visual manipulation. Specifically, this type of deepfake technology concentrates on producing highly realistic synchronization between audio and video, thus giving the impression that the person in the video genuinely speaks the words heard in the audio [13]. Developing a convincing deepfake with synchronized lip movement involves analyzing the audio for words' phonetic and timing details while using sophisticated computer vision algorithms to detect facial landmarks in the target individual's video. Through this stage, the movements of the jaw, lips, and other facial attributes are detected and precisely tracked. The phonetic and timing data obtained from the audio is aligned with the associated lip movements identified in the video using machine learning models like RNNs [162] and TCNs [187]. The lip movements are synthesized to match the timing of the audio and then combined with the original video to create a believable lip-synchronization deepfake [187]. There exist two approaches for developing talking heads that are synchronized with audio. One approach involves generating an entire individual's head as they speak, utilizing various techniques including 2D [32] or 3D [192] features and 3D meshes [31]. Nevertheless, this methodology's significant challenge is the uncertainty and imprecision related to these representations. When these methods are applied to scenarios specific to an individual [60, 136], these methods often produce poor results when used on

different people or audio sources. Neural Radiance Field (NeRF) is used in person-specific [140], and [49] modeling for talking head generation. Although this approach shows potential in generating believable talking heads for specific individuals, it has a drawback when driven by audio from another individual. The second approach is centered on synchronizing only the movement of the mouth while retaining other facial features in the video. This approach focuses on accurately animating the mouth movement to match the audio while preserving the person's identity in the video. Several methods have shown encouraging outcomes in achieving realistic lip-synchronization effects.

2.3 Audio Deepfake

In the present era, the domain of AI has made remarkable progress in the advancement of speech [75, 119], and voice-altering [186] technologies that are increasingly realistic and closely resemble human speech patterns. However, there are concerns about its misuse, one primary concern is the creation of an audio deepfake, where individuals can modify or create fake audio that appears to be from a genuine person [53]. The extensive amount of voice recordings shared on the internet has presented a challenge in detecting counterfeit audio. A growing concern is that audio deepfake attacks are no longer limited to individuals and organizations but also target high-profile politicians. The political danger presented by audio deepfake to political environments is of great concern, as it can significant unrest, confusion, and erosion of public trust [5]. To overcome these obstacles, researchers are making unceasing efforts to develop advanced and reliable methods for detecting audio content [96, 167, 174, 175]. Automated tools capable of recognizing indications of tempering, unusual artifacts, and discrepancies in audio patterns are being developed to counter the proliferation of audio deepfakes [93, 105, 171, 176, 193]. There are currently two distinct classes of audio deepfake, these classes present an ever-increasing set of challenges and difficulties regarding their identification.

- **Speech Generation**

  Speech generation, a widely recognized technological advancement often known as Text-To-Speech (TTS), is a technology that utilizes software or hardware systems to artificially generate speech that simulates human speech patterns [138].

- **Voice Conversion**

  Voice conversion (VC) is a complex process that serves to modify the speech attributes of an individual in such a way as to make it sound like the speech of another individual without modifying the original linguistic content. The primary objective of voice conversion is to manipulate a speaker's voice identity while maintaining the spoken message's integrity [186].

Tables 1, 2, 3 and 4 present a descriptive and overview of the deepfake generation techniques.

3 Literature Review

This section presents a comprehensive analysis of state-of-the-art techniques utilized for detecting and identifying audio-visual deepfakes. It offers a detailed review of detection approaches for each type of deepfake to enhance understanding of diverse methodologies from different type of review studies, such as [69, 70, 73]. The analysis critically evaluates the existing literature, encompassing technological strengths, limitations, and challenges. These deepfakes can be classified into three primary categories: facial manipulation, lip sync, and audio deepfake. The categorization was done to facilitate a better understanding of the technique employed to identify such manipulations.

3.1 Facial Manipulation Detection

Facial manipulation techniques comprise diverse methods that modify facial features or expressions in visual media. Detecting these manipulations is essential to ensure the authenticity of such content. This section provides an in-depth analysis and understanding of state-of-the-art techniques to detect facial manipulations, highlighting strengths and limitations.

3.1.1 Face Swapping Detection

Face swapping replaces a person's face in an image or video. Deep-learning algorithms are used to analyze and modify facial features for realistic transformations. The detection of this manipulation necessitates analyzing subtle inconsistencies between facial features. Traditional methods used facial landmarks and handcrafted features for comparison, but deep-learning methodologies have improved detection accuracy.

Traditional Methods

Table 1. A descriptive overview of DeepFake generation techniques.

Reference	Description
	Face Swapping Generation Techniques
Xu et al. [163]	The utilization of various loss functions during the training procedure of MobileFaceSwap has the potential to cause instability and the emergence of artifacts, owing to the absence of a paired ground truth for the constraint.
Xu et al. [164]	This framework relies on StyleGAN2, which might have limitations in preserving spatial information when transforming faces into latent vectors. The Swapping-Driven Mask branch in StyleSwap might not always achieve optimal information fusion, resulting in suboptimal face-swapping outcomes.
Kim et al. [76]	The limitation of this method is that when dealing with complex settings or visuals with obstructions like eyeglasses or partial obstructions of the face, it may encounter challenges in managing these situations, resulting in inaccurate or incomplete exchanges.
Jiang et al. [61]	Its limited ability to handle occlusions, its inability to remove glasses from the source image, and its challenges in restoring complex illuminations.
Yoo et al. [179]	The self-supervised training approach employed by FastSwap may not capture all the intricacies and variations in face attributes, thus presenting potential limitations in the realm of attribute editing.
Zhu et al. [194]	Struggle with extreme variation in facial shape and occlusions.
Yoo et al. [178]	Computational Constraints. Struggles when the source and target have significantly different facial structures.
Wang et al. [153]	Occluded or side views may result in decreased performance because of the reliance on frontal image quality.
	Face-Synthesis Generation Techniques
Yin et al. [177]	The utilization of imperfect GAN inversion for attribute editing can result in errors in reconstruction and a loss in the fidelity of textures. At the same time, the challenges in managing facial occlusion can lead to anomalies in obstructed regions. The limitations of StyleGAN give rise to issues of texture-sticking, thereby impacting the coherence of facial components in the generated videos.
Nagahara et al. [106]	The main drawback of this method lies in the compromised quality of the reconstructed frames.
Ye et al. [173]	Occasionally, occasional fluctuations may result in artifacts, such as shaking hairs, in the generated landmarks sequences.
Kim et al. [78]	The absence of 3D uniformity across different poses, and the necessity to balance label consistency and diversity.
	Attribute Manipulation Generation Techniques
Kwak et al. [111]	The linear relationship assumption utilized in OLS regression can impose certain constraints when applied to attribute fitting. The OLS performance is limited when modifying specific attributes, such as age, and can lead to unpredictable output details.
Shao et al. [131]	CCFAM demonstrates the ability to generate more disentangled edits on real faces. However, achieving complete disentanglement of attributes remains a challenging task.
Pernus et al. [117]	In comparison to encoder-decoder techniques, MaskFaceGAN operates at a slower pace. It may generate inconsistent outcomes for certain attributes.
	Puppeteering generation techniques
Thies et al. [142]	Face occlusions, such as hair and beard, and sparse mouth behavior data cause temporal aliasing issues. Real-time performance is affected by hardware-induced delays, but specialized hardware is not used for accessibility purposes.
Wang et al. [147]	Occasional distortions or unrealistic details can persist, particularly when dealing with complex expressions.
Solanki et al. [133]	Their system's accuracy relies on collecting different facial expressions from each subject to train the Expression Decoder Network. However, this limits its effectiveness when access to varied training data is restricted or unavailable for certain individuals.
Li et al. [88]	The quality of facial expressions depends on face alignment methods and is affected by complex backgrounds, necessitating method refinement and larger training datasets for better adaptation.

Zhang et al. [184] presented a methodology intending to identify swapped faces using the SURF descriptor. This local feature extraction technique extracts unique key points and patterns in images. The method

Table 2. A descriptive overview of DeepFake generation techniques - continued.

Reference	Description
	Lip-Sync generation techniques
Prajwal et al. [120]	The generative model is built on low-quality images, potentially leading to blurry output.
Park et al. [115]	Increasing the number of slots beyond a specific point could potentially have an adverse influence on performance, suggesting a compromise between the number of slots and the efficiency or effectiveness of the model.
Wu et al. [159]	Performance is strongly dependent on the training data's quality and diversity.
Guan et al. [45]	It is limited in its capacity to alter the position of the head and facial expressions due to constraints imposed by fixed masks. On certain occasions, individuals with large jaw sizes may exceed the boundaries of the mask, resulting in a compromise in visual coherence.
Zhong et al. [190]	Style aggregation quality is decreased by poor or short video reference.
Bao et al. [22]	Requires clear segmentation of the mouth and neck and is not tested in extreme poses or occlusions.
	Speech Synthesis Generation Techniques
Sarfjoo et al. [127]	The absence of proper conditioning on textual data may lead to the production of fictitious yet linguistically plausible terms that may initially appear authentic but are unnatural upon closer inspection.
Wang et al. [154]	The training process of Tacotron may necessitate significant engineering due to the complexity of modern TTS design.
Vecino et al. [144]	The enhancement in computational resources due to LE2E has the potential to undermine its overall performance.
Kang et al. [63]	The potential drawback of emotional expressiveness and speech quality could be considered as a noteworthy constraint.
Saito et al. [124]	Potential hallucinations of ChatGPT responses have the potential to exert an influence on the quality and reliability of synthesized speech.
Baas et al. [21]	ASGAN is limited to producing utterances with a predetermined length and encounters difficulty in generating cohesive complete sentences when faced with datasets containing longer utterances.
Lyth et al. [95]	1) The model struggles to control reverberation precisely.
Tan et al. [137]	Limited to English speech
	Voice Conversion Techniques
Kim et al. [77]	There is still a need to improve the processing of unfamiliar speakers and utterances. The residual encoder employed in ASSEM-VC faces challenges when it comes to accurately encoding the intonation of unfamiliar speakers, resulting in diminished MOS.
Tanaka et al. [139]	Their model still requires careful hyperparameter tuning to achieve optimal results.
Yang et al. [170]	Its performance is not consistently the best across all metrics. Their system performance is comparable to other systems such as S3PRLVC AND S2VC, showing that it is equivalent rather than much superior.
Zhang et al. [182]	The model was only evaluated on English-Spanish. GPU memory limitations resulted in the utilization of only three reference utterances.
Kumar et al. [84]	Generalization is reduced by poorly handling rare emotions and overfitting to common ones.

achieved satisfactory performance for face swap images but struggled with more complex manipulated videos. Matern et al. [99] presented an approach for deepfake detection that relies on the utilization of absent features in the region of the eye and teeth, such as missing reflections and details. However, this approach is only applicable to images that meet certain prerequisites, such as open eyes and visible teeth. Yang et al. [172] approximated the 3D orientation of the head to identify face-swapped deepfakes, but the technique's performance was compromised in cases of blurred images or variation in landmark orientations, which decreased the efficiency. Kharbat et al. [74] proposed a methodology for detecting deepfakes that utilizes the inconsistency arising from boundary effects. This method could face difficulties in detecting deepfakes that are highly complex and created using advanced techniques. Guera et al. [47] method centered on identifying AI-generated faces within the video. They used multimedia stream descriptors to extract features from individual video frames. This approach has shown effectiveness in detecting manipulated faces within the video. Despite this, the approach has faced challenges when altering

Table 3. A overview of DeepFake generation techniques

Reference	Datasets	Features	Techniques
	Face Swapping Generation Techniques		
Xu et al. [163]	FaceForensics++ [123], VGGFace2 [29]	DeepFeatures	MobileFaceSwap
Xu et al. [164]	VGGFace [116], Faceforensics++ [123], Vox-Celeb2 [134], CelebA-HQ [64]	DeepFeatures	StyleSwap
Kim et al. [76]	FFHQ, VGGFace2 [29]	DeepFeatures	Smooth-Swap
Jiang et al. [61]	FFHQ, CelebAMask-HQ [87], FaceForensics++ [123]	DeepFeatures	StyleIPSB
Yoo et al. [179]	Vox-Celeb2 [134]	DeepFeatures	FastSwap
Zhu et al. [194]	FFHQ (Flickr-Faces-HQ) [118], FaceForensics++ (FF++) [123]	DeepFeatures	StableSwap (Reversible Autoencoder (VQGAN))
Yoo et al. [178]	CelebA-HQ [64], FF++ [123]	Pose, expression, shape, and semantic details, Identity Features	InFS (Inversion-based Face Swapping)
Wang et al. [153]	CelebA-HQ [64], LFW [56]	Deep Features	NullSwap
	Face-Synthesis Generation Techniques		
Yin et al. [177]	VoxCeleb [107], HDTF [185], CelebA-HQ [64]	DeepFeatures	StyleHEAT
Nagahara et al. [106]	WFLW [158]	98 points facial landmarks	Video Coding Framework (HRNet, VSBNet & VVC encoder & decoder)
Ye et al. [173]	Lrs3-ted [11]	3D-Landmarks	GeneFace
Kim et al. [78]	CASIA-WebFace [55], LFW [56], CFP-FP [129], CPLFW [188], AgeDB [104], CALFW [189]	DeepFeatures	DCFace
	Attribute Manipulation Generation Techniques		
Kwak et al. [111]	Celeb-A [94], AffectNet [103]	DeepFeatures	IricGAN
Shao et al. [131]	Own dataset	DeepFeatures	CCFAM
Pernus et al. [117]	FFHQ, FRGC [118], XM2VTS [102], SiblingsDB-HQf [145]	DeepFeatures	MaskFaceGAN
	Puppeteering generation techniques		
Thies et al. [142]	Customized	Facial Landmarks Features	Face2Face
Wang et al. [147]	AffectNET [103]	17 AUs	U-Net-based Generator with multi-attention gate
Kumar et al. [133]	Own Dataset	DeepFeatures	DFM
Li et al. [88]	CelebA [94], Internet-Searched, CAS-PEAL [43]	SIFT, PCA-SIFT	SC-GAN
	Lip-Sync generation techniques		
Prajwal et al. [120]	LRS2 [10], LRS3 [11], LRW [150]	Mel-spectrogram representation	Wav2Lip
Park et al. [115]	LRW [150], LRS2 [10]	Lip motion features, Spectrogram	SyncTalkFace
Wu et al. [159]	LRS2 [10], VoxCeleb2 [134]	Mouth Region, DeepFeatures	Speech2Lip
Guan et al. [45]	VoxCeleb2 [134], LRW [150]	DeepFeatures	StyleSync
Zhong et al. [190]	VoxCeleb [107], HDTF [185]	13 expression parameters related to mouth motion (3DMM)	StyleSync
Bao et al. [22]	VoxCeleb2 [134], In-the-wild videos	MFCC, Facial Landmark Points, mouth area masks	Modulated Inpainting Lip-sync GAN architecture (MILG)

the video's encoded parameters.

Deep-Learning-Based Methods

Numerous research studies have employed deep learning methodologies to detect manipulated videos

Table 4. A overview of DeepFake generation techniques - continued.

Reference	Datasets	Features	Techniques
		Speech Synthesis Generation Techniques	
Sarfjoo et al. [127]	VCTK(44hrs)	Linguistic features, Frequency, Log f0	WaveNet
Wang et al. [154]	Internal North American English(24.6 hrs)	DeepFeatures	Tacotron
Vecino et al. [144]	LJSpeech (24hrs) [6]	Acoustic latent	Lightweight E2E-TTS
Kang et al. [63]	Multi-Speaker [181]	Identity and emotional features	ZET-Speech
Saito et al. [124]	Own dataset	DeepFeatures	ChatGPT-EDSS
Baas et al. [21]	SC09	DeepFeatures	ASGAN
Lyth et al. [95]	Multilingual LibriSpeech (MLS) [121], LibriTTS-R [81], EdAcc [126], VCTK [127], and VoxPopuli English subset [146]	Accent Labels, Pitch and Speaking Rate, Natural Language Metadata	Adapted AudioCraft model + DAC codec + Cross-attention
Tan et al. [137]	LJSpeech [6], VCTK [127], News-Crawl Text Corpus	Phoneme Sequences, Frame-Level Latent Representations, Duration, F0, Prosody	NaturalSpeech
		Voice Conversion Techniques	
Kim et al. [77]	VCTK [127], LibriTTS [181]	PPG, Cotatron	ASSEM-VC
Tanaka et al. [139]	Japanese speech dataset [86]	Phonetic information	PRVAE-VC
Yang et al. [170]	VCTK [127], CMU ARCTIC [170]	Mel-S3R representation	Mel-S3R VC, VQ-VAE
Zhang et al. [182]	LibriTTS [81], Multilingual LibriSpeech [121], VCTK [127], M-AILABS (Spanish)	Deep Features	RefXVC(cross-lingual voice conversion)
Kumar et al. [84]	Emo V-DB, CREMA-D [132]	Mel-spectrograms, Pitch and Energy, Phoneme Sequences, Speaker Embeddings, Emotion Embeddings, Intensity Embeddings	Emotional VC (sequence-to-sequence encoder-decoder model) with intensity Control Module and Mixed Embeddings technique

created using face-swapping techniques. David Guera and Edward J. Delp [48] presented an approach based on CNN to detect fake videos. The approach extracted features at the frame level and used RNN to detect fake videos. Despite the promising result obtained with this method, it was restricted to a short video duration of 2 seconds. Li et al. [91] used the Dlib software package to detect faces and extract face regions from the original image and then trained the four CNN models for detecting the presence of manipulated content in videos. Although this method effectively detects manipulated content, it faces difficulty in detecting videos that have undergone multiple rounds of compression. Li et al. [90] proposed a method that identified inaccurate eye blinking in manipulated faces as an indication of counterfeiting. Although this technique improves detection accuracy, its effectiveness is limited to scenarios where the lack of eye blinking is the only sign of forgery. Nguyen et al. [110] developed a convolutional neural network utilizing multitasking learning to detect manipulated images and videos and accurately identify manipulated regions. The network demonstrates its adaptability to unknown attacks through the fine-tuning process that requires only limited data, but complex manipulations may result in misclassification and inaccurate segmentations. Agarwal et al. [12] introduced a forensic technique integrating facial and behavioral biometrics to detect face-swap manipulations. Although this technique effectively detects face swap, it may not be generalized well to lip-sync-based deepfakes. Bonettini et al. [26] proposed a modified version of EfficientNetB4 with the integration of the attention mechanism. The models are trained using both the end-to-end and Siamese training paradigms. The proposed method improved the performance, but the network may be susceptible to adversarial attacks. Ismail et al. [58] proposed a new technique called YOLO-CRNNs to detect face-swapped deepfake videos using spatial and

temporal features from facial regions. The method surpasses current state-of-the-art but has limitations in keeping up with evolving deepfake techniques. Kim et al. [79] proposed FReTAL, a method based on transfer learning that leverages Representation learning (ReL) and Knowledge Distillation (KD) paradigms to detect deepfakes effectively. The limitation of this approach is that it assumes that similar features exist between different types of deepfake, which may not always hold true. Coccomini et al. [35] combined mixed convolutional-transformer networks, specifically EfficientNet B0, to extract features and vision transformers for global understanding; the method achieved exceptional outcomes without relying on distillation. Moreover, the paper presented a practical voting scheme for multi-face videos to detect manipulations. The voting scheme assumes that the face identifiers accurately represent the individual actors in the video, which may not always be true. Khan et al. [72] introduced a hybrid transformer network for deepfake detection, which demonstrates competitive outcomes. However, the model's ability to generalize to real-world scenarios requires confirmation. Huang et al. [54] present a new way to identify face swapping by focusing on implicit identity. The framework uses a CNN to embed facial images into the implicit identity space. The results are promising but there are limitations in detecting advanced face-swapping techniques and real-world applicability. Sun et al. [135] introduced FakeTracer, which adopts a proactive approach to detect face-swap deepfakes by implanting traces during training. The identification and analysis of these traces serve as an effective means of detecting and exposing face-swapped deepfakes. This approach assumes deepfake models will create noticeable traces. However, more advanced techniques may avoid these traces. The method's effectiveness depends on the quality and capabilities of deepfake models, which could lead to false results. Zhou et al. [191] presented IIN-FFD (Intra-Inter Network for Face Forgery Detection), which utilizes supervised and self-supervised learning to identify different forgeries' shared and unique characteristics. It achieved better accuracy and generalization to unknown forgeries; however, it struggles in extracting similar features across diverse forgery datasets with non-overlapping artifacts. Nawaz et al. [109] utilized a modified Inception-Swish-ResNet-v2 CNN for spatial feature extraction and a Bi-LSTM for temporal features. Grad-CAM-based explainability helped it achieve great accuracy and robustness. However, the performance dropped in Cross-Dataset testing.

3.1.2 Face Generation Detection

Traditional Methods

Traditional methods have played a crucial role in detecting and mitigating the potential risk associated with face generation. McCloskey et al. [101] drew attention to the evident differences in color treatment between images originating from camera sensors versus those that GANs [44] create. The study focused on extracting color characteristics and using SVM to differentiate between authentic and artificially created images. However, this method has limitations when facing compression and larger training datasets. Guarnera et al. [46] introduced an EM algorithm-based approach using key points as image markers, and trained classification models like K-nearest neighbors, SVM, and LDA. The technique shows potential in detecting AI-generated images, but compressed images may impact its accuracy. Yang et al. [172] proposed a technique to detect GAN-generated images by carefully analyzing facial landmark points. The technique achieved impressive accuracy. However, the effectiveness of this technique may be decreased when dealing with advanced GAN models. Matern et al. [99] used eye color differences, extracting color features and computing dissimilarity measures. This method relies on the assumption of consistent differences in eye color for generated faces, which may not hold true for modern generating techniques.

Deep-Learning-Based Methods

The implementation of deep learning methods has brought a significant transformation in the field of face synthesis detection. Dang et al. [36] introduced the CGFace model to distinguish computer-generated faces from real ones utilizing customized Convolutional Neural Network (CNN) architecture. The model's capability to accurately identify images from novel and unfamiliar GAN models, which were not encountered during the training phase, may be limited. Nataraj et al. [108] proposed a method combining co-occurrence matrices and deep learning to detect GAN-generated fake images, showing its effectiveness on different GAN datasets. However, the method's performance is reduced in categories with JPEG compression. Barni et al. [23] presented a method using CNNs to distinguish between GANs [44] and genuine images, focusing on synthetic face images. The method exploits inconsistencies between spectral bands and performs better, but it might be susceptible to intentional attacks.

Chen et al. [30] introduced a novel perspective by focusing on detecting localized GAN-generated faces, utilizing an improved Xception model. Unlike previous research on entire faces, the current work targets the intricate task of identifying smaller, locally generated regions. The limitation of the study is that it focuses on image-level detection rather than the more complex problem of pixel-level localization. Wang et al. [148] presented a semantic-based approach between genuine and GAN-generated face images by thoroughly analyzing inter-eye symmetries and inconsistencies. The results showed that the proposed method could obtain equivalent or even better results than state-of-the-art techniques considering the whole face image. However, it relies on the assumption of GAN-generated inter-eye inconsistencies, which may become invalid as GAN technology advances. Mandelli et al. [98] proposed a new method that employs ensemble CNNs. The proposed approach employs the diversity of orthogonal training datasets to enhance generalization to unseen image generators. The method employs patch-level analysis and assigns an overall image label based on the presence of synthetic patches and their scores. Although this may provide robustness against local changes, it might struggle to capture the global image context. The detection accuracy of the method may face difficulties when dealing with synthetic images that do not possess distinguishable patch-level characteristics. Xue et al. [165] introduced GLFNet, a novel deepfake detection method that combines physiological attributes with deep learning techniques. The method includes a local region branch and a global detection branch. While the method shows promising results, its vulnerabilities to adversarial attacks are not explored. Boyd et al. [27] presented the CYBORG approach, which aims to improve the generalization abilities of deep learning models by integrating human perceptual insights. The strategy directs the model's learning process to focus on parts of images that humans find visually important by adding human-annotated saliency maps into the model's loss function. Biases and inconsistencies may arise if human accuracy is not at an expert level.

3.1.3 Attributes Manipulation Detection

Traditional Methods

These methodologies utilize manually crafted features to detect manipulated facial attributes. Scherhag et al. [128] proposed a PRNU-based detection system to distinguish genuine from morphed facial images. They examined the spatial and spectral characteristics of PRNU patterns affected by facial warping. The proposed approach involves preprocessing, PRNU extraction, feature extraction, feature aggregation, and decision-making. The system's accuracy could be reduced due to the scanner's PRNU. This limitation highlights the difficulties in identifying morphed images in printed and scanned media scenarios. Li et al. [183] proposed a technique for identifying GAN-generated images without particular GAN models. The paper proposed two methodologies: identifying artifacts produced by GAN up-sampling and creating a classifier using frequency spectra rather than pixels and using a GAN simulator called AutoGAN. The methodology's performance was successful with StarGAN, while a significant drop was observed with GauGAN.

Deep-Learning-Based Methods

Due to the pervasive impact of deep learning, numerous studies have focused on utilizing Deep learning models to identify attribute manipulation. For instance, Wang et al. [152] introduced FakeSpotter, a new approach for observing deep face recognition systems using neuron coverage techniques. The approach captures subtle features that distinguish genuine from fake faces by analyzing layer-by-layer neuron activation patterns. Mean Neuron Coverage (MNC) is introduced as a criterion to identify significant differences between genuine and synthesized faces. The method's effectiveness in detecting samples decreased notably when faced with lighting variation. Akhtar et al. [15] created a deepfake dataset and evaluated deep learning features for identifying deepfakes. Deep feature-based detection systems perform well on the same manipulation type but struggle with novel types. Yu et al. [180] proposed a framework for detecting manipulated fake faces generated by AI. The framework mined intrinsic features from channel difference and spectrum images, eliminated bias among manipulation techniques, and achieved competitive performance. However, it struggles with new types of manipulation. Guo et al. [50] proposed GocNet, which involves two plug-and-play modules, tensor preprocessing (TP) and manipulation trace attention (MTA), that can enhance the performance of CNNs when detecting face forgeries. Both modules can be easily integrated into existing CNNs for training and have outperformed prior works on five public image datasets. However, the computational intensity of GocNet with TP and MTA modules may limit its deployment on constrained platforms. Wang et al.

[149] proposed a technique for categorizing edited facial attributes by GANs [44] using patch-based hybrid classification networks with localization supervision. The process operates in two phases: splitting the image into patches for analysis and integrating the features obtained from these patches. The method shows accurate and robust results on 19 facial attributes modified by StyleGAN2 [66]. However, the generalization capability of the method may have limitations for unseen manipulations.

3.1.4 Puppeteering

Traditional Methods

Traditional methods, which are based on handcrafted features, have focused on extracting and analyzing predetermined attributes of the face. For instance, Agarwal et al. [13] proposed a method that analyzes the video's facial and head movements using the OpenFace2 toolkit. The study extracted 17 action units (Aus) and additional features to capture distinct expressions. The Pearson correlation between these features is then measured to characterize individual motion signatures. This correlation-based approach forms feature vectors for 10-second video clips, which are then employed with SVM for detection. This approach has a limitation as it is sensitive to different speaking contexts. Amerini et al. [18] presented a forensic technique for differentiating between fake and original video sequences by using an optical flow field as the primary feature. The study used PWC-NET to compute optical flow fields and used it as input to train deep learning models. The technique showed better performance, but only preliminary results were reported and require further testing to confirm reliability.

Deep-Learning-Based Methods

Deep learning has been revolutionary in computer vision, particularly in image analysis. Pattern recognition and data-driven understanding. It offers a promising solution for detecting expression manipulation, capturing even the most subtle visual cues that traditional methods may miss. Early works, such as Afchar et al. [9] used a mesoscopic approach to focus on the properties of images that are larger than pixels but smaller than the whole image. The study employed two deep learning models, Meso-4 and MesoInception-4, for feature extraction and classification. This technique involves assessing visual features and patterns. It may face difficulties with low-quality inputs, reducing the accuracy of detection. Rana et al. [122] presented the DeepfakeStack technique involving two main steps: Base-Learners Creation and Stack generalization. DFC is trained on a holdout test dataset with updates limited to new hidden and output layer weights. Each base-learner's output is input to DFC, and predictions are combined through concatenation. This technique showed excellent performance, but the real-time performance of the technique may be limited due to its computational complexity, especially on resource-constrained devices. Kumar et al. [85] introduced a deep-learning-based system involving five parallel ResNet-18 models designed to detect reenacted frames. This multi-stream network utilizes RGB frames to capture localized facial artifacts and noise patterns introduced during the reenactment. The technique detects regional artifacts using multiple parallel streams, resulting in higher computational complexity. Wodajo et al. [156] proposed a Convolutional Vision Transformer (CViT) for detecting deepfake videos. The CViT model combined a CNN and ViT and focused on learning image features through CNN and Transformer. The CViT model is trained on the DFDC [42] datasets, which may limit its generalizability to other datasets and real-world scenarios. Mazaheri et al. [100] proposed a two-stream network for Facial Expression Manipulation Detection (EMD). The method achieved higher performance in forgery detection and localization of manipulated regions, with better classification and localization. However, the method's effectiveness depends on the accuracy of the FER model. Waseem et al. [155] proposed a novel technique utilizing an attention-based multitask framework to improve feature maps for classification and localization. The network merged frequency domain and spatial features using bilinear pooling. The network achieved superior detection performance due to focusing on forgery regions rather than biases and artifacts, resulting in more precise predictions. However, it may still struggle to detect small forgeries in low-quality videos. Liang et al. [92] presented a new method that used FGPM to extract facial geometry features and combine these features with upscaled feature maps generated by a CNN-LSTM network, which captures spatial and temporal information. The model successfully located manipulations at the pixel level. However, the model's performance is limited to the specific manipulations it was trained on, it may not detect new and evolving techniques.

3.2 Lip-Sync Detection

Traditional Methods

Traditionally, lip sync detection has depended upon the spectrum of handcrafted methods that analyze various aspects of video and audio content. One of the techniques proposed by Korshunov et al. [83] utilized MFCCs having different components, including the main MFCCs, their deltas, double-deltas, and energy features. The study trained multiple classifiers including SVM, LSTM, MLP, and GMM. LSTM outperformed, but its accuracy dropped when tested with datasets such as VidTIMIT and AMI. Another technique proposed by Jafar et al. [59] utilized the mouth as a feature and cut the video based on specific words in which the mouth appears open, and teeth are visible. They presented a deepfake detection model with mouth features (DFT-MF). The model achieved high accuracy on Celeb-DF and Deepfake Vid-TIMIT datasets. The model heavily depends on the mouth region. If deepfake videos avoid mouth artifacts, the model's accuracy may decrease.

Deep-Learning-Based Methods

Deep learning techniques have transformed the field of multimedia forensics. They offer solutions for detecting lip sync deepfakes, which are getting harder to detect. Many research studies have focused on using deep learning models to identify and detect lip sync deepfake. For instance, Chintha et al. [33] introduced a framework called XcepTemporal convolutional recurrent neural network to detect deepfakes. The author utilized an XceptionNet CNN as a facial feature representation, which is passed through bidirectional recurrence layers to detect temporal inconsistencies. The study also introduced a complementary architecture for audio spoofing. The approach's effectiveness was reduced when employed on compressed samples, thus revealing a constraint in its ability to process compressed data. Haliassos et al. [51] presented a LipForensics approach that utilizes CNN that is pre-trained on lipreading to acquire semantic representations of mouth dynamics. The approach is limited to accurately identifying fake videos where the mouth is occluded or unmodified. Furthermore, its performance may also decline in scenarios with restricted mouth movements. Ilyas et al. [57] proposed a hybrid deep learning framework called InceptionResNet-BiLSTM using a customized InceptionResNetV2 and a Bidirectional LSTM. The customized InceptionResNetV2 is specifically designed to extract frame-level learnable features from faces extracted from the videos. These features are then arranged in sequences and utilized to train temporally aware BiLSTM to distinguish between genuine and fake videos. The model encounters difficulties in adjusting to various deepfake techniques, affecting its consistency. Its high resource requirements might limit widespread deployment, especially in resource-constrained environments. Shahzad et al. [130] presented the "Lip Sync Matters" method that targets high-level semantic features. The method utilized the ResNet-18 model to identify lip movements in the video, and a Wave2lip model was used to create a synthetic lip sequence from the audio. The method depends on visible frontal faces and is susceptible to adversarial attacks. Agarwal et al. [14] presented a multi-model semantic forensic approach for classifying videos as genuine or fake based on the alignment of facial expressions with speech. The method uses interpretable action units (AUs) to capture a person's face and head movement. Facial feature vectors are created for each word to capture motion during pronunciation. Linear word-specific classifiers are trained to detect whether facial movements match spoken words. The method evaluates video features using these classifiers and computes a final authenticity score. However, the method has limitations due to requiring person-specific training data and may be less reliable for unconstrained videos. It has only been validated for English speech and does not address video manipulation artifacts, making it vulnerable to deepfake techniques. Yang et al. [171] introduced a new method called AVoiD-DF to identify deepfake videos using audio and visual elements. They used a Temporal-Spatial Encoder (TSE) to capture inconsistencies in temporal-spatial information, a Multi-Model Joint-Decoder (MMD) to fuse audio-visual features, and a Cross-Model Classifier to detect deepfake manipulations. The method outperformed existing techniques, but its performance relies on high-quality audio-visual training data. Bohacek and Farid [25] employed Whisper and Auto-AVSR models to compare audio and video transcriptions and measure the mismatch between them using normalized Levenshtein distance. However, it achieved great accuracy; however, very short videos had lower detection accuracy due to the limited context. Datta et al. [38] proposed LIPINC to identify lip-sync using spatial and temporal inconsistencies in the mouth region. It extracts mouth frames and processes the variation in color and structure using 3D-CNN and cross-attention. LIPINC demonstrated great accuracy and generalization but struggled with short videos and lacked analysis of audio-visual mismatch. To address this, Datta et al.

[39] utilized Vision Temporal Transformer (VTT) to examine spatial and temporal inconsistencies in mouth motions. It allows segment-wise localization of modified regions using multi-head cross-attention. Although it performs well on multiple datasets, it struggles to detect global frames in short videos.

3.3 Audio Deepfake

Detecting audio deepfakes is a crucial area of study involving various techniques to authenticate voice recordings. These techniques comprise voice analysis, prosody analysis, emotion analysis, and prosody analysis for speech generation (TTS) deepfakes, while speaker verification, behavioral analysis, and acoustics for voice conversion deepfakes. The advancement in AI and ML plays a pivotal role in evolving these detection methods. However, the challenge lies in maintaining pace with the continuously advancing sophistication of audio deepfake technologies. This necessitates ongoing research and efforts to enhance accuracy and effectiveness in preserving the authenticity of audio content. This section has provided an overview of the methods utilized for detecting both types of audio deepfakes.

3.3.1 Speech Generation

Traditional Methods

Pal et al. [113] presented a technique to identify synthetic speech by focusing on three categories of artifacts linked to magnitude, phase, and pitch variation. They employed a score-level fusion of diverse front-end features and used the GMM-based classifier. The method achieved outstanding performance in detecting spoofing. The ASVspoof2015 [161] database used to train the synthetic speech detector is limited in its scope of spoofing attacks, limiting the generalization ability of the countermeasures systems. Yang et al. [168] proposed an extended CQCC feature, combining information from linear and octave power spectra of speech signals. They employed the DNN classifier as the back-end system. The method is effective for synthetic and replay spoofing attacks but not for other attacks like voice conversion or impersonation. This limits its generalizability. Li et al. [89] proposed a Res2Net to identify audio manipulation. They used various acoustic features and found that Res2Net performs optimally when Constant-Q Transform (CQT) features are employed. Although their method performs well, it may be susceptible to adversarial attack, and the ability of the model to generalize in different situations requires improvement. Another approach proposed by Yang et al. [169] presented three features based on subband transforms, intending to capture distinct synthetic speech attributes and artifacts. Extensive experimental evaluations were performed on the subband features in conjunction with traditional, utilizing the ASVspoof2015 [161] and ASVspoof2019 [143] datasets. They evaluated these features under various noises such as Volvo, babble, café, and street. However, it may not cover real-world noise conditions.

Deep-Learning-based Methods

In speech synthesis detection, there has been a significant advancement in technology due to the groundbreaking work of researchers utilizing deep learning. Wu et al. [160] proposed FG-LCNN, proposed FG-LCNN which used a genuinization transformer to learn the characteristics of real speech. The proposed method performs better than baseline LCNN methods, particularly on the evaluation set with unseen spoofing attacks. Still, it heavily depends on a large and diverse dataset of genuine speech samples. The system's performance may suffer if the training data is limited or not representative of all variations. Pal et al. [114] suggested a way to tackle the problem of generalizing spoofing detection by introducing prototypical loss in the meta-learning framework. Despite efforts to increase generalization, the method may have difficulty with new or unconventional spoofing attacks that were not adequately represented in the training data. Salvi et al. [125] proposed an approach that aims to decrease the computational time needed for analyzing extended audio tracks by addressing the problem of synthetic speech detection. Their method involves folding the audio tracks multiple times, resulting in shorter tracks containing overlapping speakers. Folding the audio track multiple times is a form of data augmentation. However, this method may not work well for longer audio tracks or those with complex overlapping speech. Khan et al. [67] presented SpoTNet, an innovative approach for effectively identifying synthetic speech in the context of voice spoofing attacks. The approach introduced a distinctive element known as LSTE that extracted attentive characteristics from speech signals and combined traditional and deep learning-based methods to accurately extract acoustic features that captured the distinctive characteristics of genuine and synthetic speech samples. The method performed well but heavily relies on a large amount of training data. Yadav et al. [166] proposed a technique called ASSD to identify artificial speech using data

derived from the AAC compressed bitstream. ACC compressed signals are denoted as a series of blocks, and a feature vector with N dimensions is extracted for each block. A Transformer neural network is employed to process the feature vectors and acquire a representation of each block. The method showed effectiveness in compression to existing methods. Still, it is specifically designed for AAC compression and might not be directly suitable for alternate audio compression formats such as MP3 or FLAC.

3.3.2 Voice Conversion

Traditional Methods

Das et al. [37] utilized long-range acoustic features, specifically those obtained from the CQT, to detect and record spoofing artifacts. The study categorized these features into full band, subband, and hybrid features. These features are then combined with GMM and DNN to form models. The method showed better performance, but the effectiveness of their method heavily depends on the availability and quality of training data. Aljasem et al. [17] presented the SASV approach, which utilizes Sm-ALTP features to capture voice-specific characteristics and attack-related artifacts. These features use adaptable thresholds and collect vocal tract information, improving speaker identification and recognition abilities. The system is computationally efficient, but its performance relies on the specific attack algorithms and might need new methods updates.

Deep-Learning-based Methods

Wang et al. [151] proposed DeepSoner, a method using two criteria for neuron coverage: CAN and TKAN, to classify behaviors of neurons at different layers as features for classification. A shallow neural network is used as the binary classifier, utilizing vectorized representations of neuron behaviors as inputs. The method is accurate and robust but vulnerable to adversarial attacks. In real-world settings with high-intensity noises, the model's effectiveness may decrease. Ma et al. [96] presented a continual learning-based approach called DFWF that integrated LwF and PSA. It retains knowledge of the original model while training with new data to prevent catastrophic forgetting. Additionally, PSA is utilized to align the distribution of genuine speech features across different scenarios. The method outperformed fine-tuning when new spoofing attacks were introduced but did not outperform multi-condition training. Wu et al. [157] used a modified SENet model with a squeeze-and-excitation network and a question-answering strategy to train the model to predict the start and end positions of fake clips in the audio. A self-attention mechanism is added to improve the model's ability to focus on relevant regions. The study also utilized data augmentation techniques and experiments with different input features for better robustness. It achieved competitive performance through model fusion, but the self-attention mechanism and complex model architectures can be expensive. This might limit scalability and applicability in resource-constrained environments. Dawood et al. [40] presented CLS-LBP, a feature representation method capturing dynamic speech characteristics of authentic audio, the artifacts produced by cloning algorithms, and the alterations in replayed signals caused by microphone distortions. The system's capability to recognize particular cloning algorithms depends on previous knowledge of these algorithms. It may face difficulties in identifying spoofing attempts that utilize unknown techniques. Deng et al. [41] proposed a technique called REVELIO, specifically designed to restore voiceprints from audios that have undergone voice conversion techniques. The method used representation learning to extract the original speaker's voiceprint. The study also used evidence audio to refine the extracted voiceprint. The method was effective across various voice conversion techniques. However, its performance can be influenced by the adaptability of voice conversion processes, challenges in inverse voice conversion, and the quality of the original voice conversion. Bird and Lotfi [24] proposed a method to identify real-time AI-generated voice conversion (RVC) attacks by utilizing machine learning techniques. The process involves gathering audio data and extracting various attributes, such as spectral characteristics and MFCCs, from real and AI-generated datasets. Statistical analysis and machine learning models were used to detect disparities between the two types of data. The method relies on the DEEP-VOICE dataset, which may have limitations. Ma et al. [97] proposed SDDE (self-distillation-based Domain Exploration) that integrates supervised learning and self-distillation in a shared encoder framework. The model used multi-scale speech segments, and Performance is enhanced with longer speech segments (up to 14 seconds), but training time and computational costs are significantly increased.

The limitations associated with current state-of-the-art methods in the detection of deepfakes encompass several critical challenges. These techniques face

challenges in applying them to different deepfake techniques due to the significant variability in the manipulation approaches used and the lack of comprehensive training data. Additionally, these methods are also vulnerable to adversarial attacks and have limited scalability. Addressing these limitations necessitates a collaborative effort to enhance the ability to generalize across diverse manipulation techniques, expand and diversify training datasets, strengthen detection systems against adversarial tactics, and optimize scalability for broader deployment across various environments.

4 Conclusions

The proliferation of deepfake technologies has introduced an unprecedented level of realism in manipulated multimedia content, posing significant risks to privacy, security, and information authenticity. This survey comprehensively examines the landscape of deepfake generation and detection, highlighting the rapid evolution of AI-driven manipulation techniques and the corresponding countermeasures. Deepfake generation methods, such as face swapping, attribute manipulation, and audio synthesis, leverage advanced models like GANs and VAEs to produce highly realistic forgeries, raising concerns about misinformation, privacy violations, and political manipulation.

Detection approaches have similarly advanced, with traditional methods relying on spectral and behavioral inconsistencies, while deep learning-based techniques exploit spatial-temporal features and multimodal analysis. However, significant challenges persist: (1) Generalization: Many detectors fail against unseen manipulation techniques or cross-dataset scenarios. (2) Adversarial Robustness: Deepfake generators can bypass detectors using adversarial attacks. (3) Scalability: Real-time detection remains computationally demanding, especially for high-resolution videos.

The paper underscores the importance of standardized benchmarks (e.g., DFDC, ASVspoof) and interdisciplinary collaboration to address these gaps. Future research should focus on:

- Explainable AI: Developing interpretable models to enhance trust and debuggability.
- Multimodal Fusion: Integrating audio, visual, and behavioral cues for holistic detection.
- Continual Learning: Adapting detectors to emerging deepfake variants through dynamic training.
- Ethical Frameworks: Establishing guidelines for responsible deepfake use and mitigation.

In conclusion, while deepfake technology poses formidable challenges, ongoing innovations in detection methodologies offer promising avenues to safeguard digital media integrity. This survey provides a foundation for researchers to advance robust, scalable solutions and foster public awareness of deepfake risks.

Data Availability Statement

Not applicable.

Funding

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] FaceApp: Face Editor. Retrieved from https://www.faceapp.com/

[2] FakeApp. (2019, March 7). Malavida. Retrieved from https://www.malavida.com/en/soft/fakeapp/

[3] Deepfakes/faceswap: Deepfakes software for all. (n.d.). Retrieved from GitHub.https://github.com/deepfakes/faceswap

[4] Iperov/DeepFaceLab: DeepFaceLab is the leading software for creating deepfakes. (n.d.). GitHub. Retrieved from https://github.com/iperov/DeepFaceLab

[5] Ratnam, G. (2020, March 2). How fake audio, like deepfakes, could plague business, politics. Roll Call. Retrieved from https://rollcall.com/2020/03/03/how-fake-audio-like-deepfakes-could-plague-business-politics/

[6] The LJ speech dataset. (n.d.). Kaggle: Your Machine Learning and Data Science Community. Retrieved from https://www.kaggle.com/datasets/mathurinache/the-lj-speech-dataset.

[7] (n.d.). Reface. Retrieved from https://reface.ai/

[8] ZAO Yunifood apps on the app store. (n.d.). App Store. Retrieved from https://apps.apple.com/by/developer/zao-yunifood/id1450822231

[9] Afchar, D., Nozick, V., Yamagishi, J., & Echizen, I. (2018, December). Mesonet: a compact facial video forgery detection network. In *2018 IEEE international*

workshop on information forensics and security (WIFS) (pp. 1-7). IEEE. [Crossref]

[10] Afouras, T., Chung, J. S., Senior, A., Vinyals, O., & Zisserman, A. (2018). Deep audio-visual speech recognition. *IEEE transactions on pattern analysis and machine intelligence, 44*(12), 8717-8727. [Crossref]

[11] Afouras, T., Chung, J. S., & Zisserman, A. (2018). LRS3-TED: a large-scale dataset for visual speech recognition. *arXiv preprint arXiv:1809.00496*. [Crossref]

[12] Agarwal, S., Farid, H., El-Gaaly, T., & Lim, S. N. (2020, December). Detecting deep-fake videos from appearance and behavior. In *2020 IEEE international workshop on information forensics and security (WIFS)* (pp. 1-6). IEEE. [Crossref]

[13] Agarwal, S., Farid, H., Gu, Y., He, M., Nagano, K., & Li, H. (2019, June). Protecting world leaders against deep fakes. In *CVPR workshops* (Vol. 1, No. 38).

[14] Agarwal, S., Hu, L., Ng, E., Darrell, T., Li, H., & Rohrbach, A. (2023). Watch those words: Video falsification detection using word-conditioned facial motion. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 4710-4719). [Crossref]

[15] Akhtar, Z., Mouree, M. R., & Dasgupta, D. (2020, September). Utility of deep learning features for facial attributes manipulation detection. In *2020 IEEE International Conference on Humanized Computing and Communication with Artificial Intelligence (HCCAI)* (pp. 55-60). IEEE. [Crossref]

[16] Albahar, M., & Almalki, J. (2019). Deepfakes: Threats and countermeasures systematic review. *Journal of Theoretical and Applied Information Technology, 97*(22), 3242-3250.

[17] Aljasem, M., Irtaza, A., Malik, H., Saba, N., Javed, A., Malik, K. M., & Meharmohammadi, M. (2021). Secure automatic speaker verification (SASV) system through sm-ALTP features and asymmetric bagging. *IEEE Transactions on Information Forensics and Security, 16*, 3524-3537. [Crossref]

[18] Amerini, I., Galteri, L., Caldelli, R., & Del Bimbo, A. (2019). Deepfake video detection through optical flow based cnn. In *Proceedings of the IEEE/CVF international conference on computer vision workshops* (pp. 0-0). [Crossref]

[19] Arik, S., Chen, J., Peng, K., Ping, W., & Zhou, Y. (2018). Neural voice cloning with a few samples. *Advances in neural information processing systems, 31*. [Crossref]

[20] Arık, S. Ö., Chrzanowski, M., Coates, A., Diamos, G., Gibiansky, A., Kang, Y., ... & Shoeybi, M. (2017, July). Deep voice: Real-time neural text-to-speech. In *International conference on machine learning* (pp. 195-204). PMLR. [Crossref]

[21] Baas, M., & Kamper, H. (2024). Disentanglement in a GAN for unconditional speech synthesis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 32*, 1324-1335. [Crossref]

[22] Bao, H., Zhang, X., Wang, Q., Liang, K., Wang, Z., Ji, S., & Chen, W. (2024). MILG: Realistic lip-sync video generation with audio-modulated image inpainting. *Visual Informatics, 8*(3), 71-81. [Crossref]

[23] Barni, M., Kallas, K., Nowroozi, E., & Tondi, B. (2020, December). CNN detection of GAN-generated face images based on cross-band co-occurrences analysis. In *2020 IEEE international workshop on information forensics and security (WIFS)* (pp. 1-6). IEEE. [Crossref]

[24] Bird, J. J., & Lotfi, A. (2023). Real-time detection of ai-generated speech for deepfake voice conversion. *arXiv preprint arXiv:2308.12734*. [Crossref]

[25] Bohacek, M., & Farid, H. (2024). Lost in Translation: Lip-Sync Deepfake Detection from Audio-Video Mismatch. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 4315-4323). [Crossref]

[26] Bonettini, N., Cannas, E. D., Mandelli, S., Bondi, L., Bestagini, P., & Tubaro, S. (2021, January). Video face manipulation detection through ensemble of cnns. In *2020 25th international conference on pattern recognition (ICPR)* (pp. 5012-5019). IEEE. [Crossref]

[27] Boyd, A., Tinsley, P., Bowyer, K. W., & Czajka, A. (2023). Cyborg: Blending human saliency into the loss improves deep learning-based synthetic face detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 6108-6117). [Crossref]

[28] Peele, J. (2018). You won't believe what obama says in this video. *Youtube*.

[29] Cao, Q., Shen, L., Xie, W., Parkhi, O. M., & Zisserman, A. (2018, May). Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)* (pp. 67-74). IEEE. [Crossref]

[30] Chen, B., Ju, X., Xiao, B., Ding, W., Zheng, Y., & de Albuquerque, V. H. C. (2021). Locally GAN-generated face detection based on an improved Xception. *Information Sciences, 572*, 16-28. [Crossref]

[31] Chen, L., Cui, G., Kou, Z., Zheng, H., & Xu, C. (2020). What comprises a good talking-head video generation?: A survey and benchmark. *arXiv preprint arXiv:2005.03201*. [Crossref]

[32] Chen, L., Maddox, R. K., Duan, Z., & Xu, C. (2019, June). Hierarchical Cross-Modal Talking Face Generation With Dynamic Pixel-Wise Loss. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 7824-7833). IEEE. [Crossref]

[33] Chintha, A., Thai, B., Sohrawardi, S. J., Bhatt, K., Hickerson, A., Wright, M., & Ptucha, R. (2020). Recurrent convolutional structures for audio spoof and video deepfake detection. *IEEE Journal of Selected Topics in Signal Processing, 14*(5), 1024-1037. [Crossref]

[34] Choi, Y., Choi, M., Kim, M., Ha, J. W., Kim, S., & Choo, J. (2018). Stargan: Unified generative adversarial networks for multi-domain image-to-image

translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8789-8797). [Crossref]

[35] Coccomini, D. A., Messina, N., Gennaro, C., & Falchi, F. (2022, May). Combining efficientnet and vision transformers for video deepfake detection. In *International conference on image analysis and processing* (pp. 219-229). Cham: Springer International Publishing. [Crossref]

[36] Dang, L. M., Hassan, S. I., Im, S., Lee, J., Lee, S., & Moon, H. (2018). Deep learning based computer generated face identification using convolutional neural network. *Applied Sciences, 8*(12), 2610. [Crossref]

[37] Das, R. K., Yang, J., & Li, H. (2019). Long Range Acoustic Features for Spoofed Speech Detection. In *Proc. Interspeech 2019* (pp. 1058-1062). [Crossref]

[38] Datta, S. K., Jia, S., & Lyu, S. (2024, July). Exposing lip-syncing deepfakes from mouth inconsistencies. In *2024 IEEE International Conference on Multimedia and Expo (ICME)* (pp. 1-6). IEEE. [Crossref]

[39] Datta, S. K., Jia, S., & Lyu, S. (2025). Detecting Lip-Syncing Deepfakes: Vision Temporal Transformer for Analyzing Mouth Inconsistencies. *arXiv preprint arXiv:2504.01470*. [Crossref]

[40] Dawood, H., Saleem, S., Hassan, F., & Javed, A. (2022). A robust voice spoofing detection system using novel CLS-LBP features and LSTM. *Journal of King Saud University-Computer and Information Sciences, 34*(9), 7300-7312. [Crossref]

[41] Deng, J., Chen, Y., Zhong, Y., Miao, Q., Gong, X., & Xu, W. (2023). Catch you and i can: Revealing source voiceprint against voice conversion. In *32nd USENIX Security Symposium (USENIX Security 23)* (pp. 5163-5180). [Crossref]

[42] Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., & Ferrer, C. C. (2020). The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397*. [Crossref]

[43] Gao, W., Cao, B., Shan, S., Chen, X., Zhou, D., Zhang, X., & Zhao, D. (2007). The CAS-PEAL large-scale Chinese face database and baseline evaluations. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 38*(1), 149-161. [Crossref]

[44] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM, 63*(11), 139-144. [Crossref]

[45] Guan, J., Zhang, Z., Zhou, H., Hu, T., Wang, K., He, D., ... & Wang, J. (2023). Stylesync: High-fidelity generalized and personalized lip sync in style-based generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 1505-1515). [Crossref]

[46] Guarnera, L., Giudice, O., & Battiato, S. (2020). Deepfake detection by analyzing convolutional traces. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops* (pp. 666-667). [Crossref]

[47] Güera, D., Baireddy, S., Bestagini, P., Tubaro, S., & Delp, E. J. (2019). We need no pixels: Video manipulation detection using stream descriptors. *arXiv preprint arXiv:1906.08743*. [Crossref]

[48] Güera, D., & Delp, E. J. (2018, November). Deepfake video detection using recurrent neural networks. In *2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS)* (pp. 1-6). IEEE. [Crossref]

[49] Guo, Y., Chen, K., Liang, S., Liu, Y. J., Bao, H., & Zhang, J. (2021). Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 5784-5794). [Crossref]

[50] Guo, Z., Yang, G., Zhang, D., & Xia, M. (2023). Rethinking gradient operator for exposing AI-enabled face forgeries. *Expert Systems with Applications, 215*, 119361. [Crossref]

[51] Haliassos, A., Vougioukas, K., Petridis, S., & Pantic, M. (2021, June). Lips Don't Lie: A Generalisable and Robust Approach to Face Forgery Detection. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 5037-5047). IEEE. [Crossref]

[52] Harwell, D. (2019). An artificial-intelligence first: Voice-mimicking software reportedly used in a major theft. *Washington Post, 4*.

[53] Harwell, D. (2021). Remember the 'deepfake cheerleader mom'? Prosecutors now admit they can't prove fake-video claims. *March, 14*, 2021.

[54] Huang, B., Wang, Z., Yang, J., Ai, J., Zou, Q., Wang, Q., & Ye, D. (2023, June). Implicit Identity Driven Deepfake Face Swapping Detection. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 4490-4499). IEEE. [Crossref]

[55] Huang, G. B., Mattar, M. A., Lee, H., & Learned-Miller, E. (2012, December). Learning to align from scratch. In *Proceedings of the 26th International Conference on Neural Information Processing Systems-Volume 1* (pp. 764-772). [Crossref]

[56] Huang, G. B., Mattar, M., Berg, T., & Learned-Miller, E. (2008). Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*.

[57] Ilyas, H., Irtaza, A., Javed, A., & Malik, K. M. (2022, December). Deepfakes examiner: An end-to-end deep learning model for deepfakes videos detection. In *2022 16th international conference on open source systems and technologies (ICOSST)* (pp. 1-6). IEEE. [Crossref]

[58] Ismail, A., Elpeltagy, M., Zaki, M., & ElDahshan, K. A. (2021). Deepfake video detection: YOLO-Face convolution recurrent approach. *PeerJ Computer Science, 7*, e730. [Crossref]

[59] Jafar, M. T., Ababneh, M., Al-Zoube, M., & Elhassan, A. (2020, April). Forensics and analysis of deepfake videos. In *2020 11th international conference on information and communication systems (ICICS)* (pp. 053-058). IEEE. [Crossref]

[60] Ji, X., Zhou, H., Wang, K., Wu, W., Loy, C. C., Cao, X., & Xu, F. (2021). Audio-driven emotional video portraits. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 14080-14089). [Crossref]

[61] Jiang, D., Song, D., Tong, R., & Tang, M. (2023, June). StyleIPSB: Identity-Preserving Semantic Basis of StyleGAN for High Fidelity Face Swapping. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 352-361). IEEE. [Crossref]

[62] Johnson, D., & Johnson, A. (2023). What are deepfakes? How fake AI-powered audio and video warps our perception of reality. *How fake AI-powered audio and video warps our perception of reality*.

[63] Kang, M., Han, W., Hwang, S. J., & Yang, E. (2023). ZET-Speech: Zero-shot adaptive Emotion-controllable Text-to-Speech Synthesis with Diffusion and Style-based Models. In *Proc. Interspeech 2023* (pp. 4339-4343). [Crossref]

[64] Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2017). Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*. [Crossref]

[65] Karras, T., Laine, S., & Aila, T. (2019, June). A Style-Based Generator Architecture for Generative Adversarial Networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 4396-4405). IEEE. [Crossref]

[66] Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8110-8119). [Crossref]

[67] Khan, A., & Malik, K. M. (2023, June). Spotnet: A spoofing-aware transformer network for effective synthetic speech detection. In *Proceedings of the 2nd ACM International Workshop on Multimedia AI against Disinformation* (pp. 10-18). [Crossref]

[68] Nightingale, S. J., & Wade, K. A. (2022). Identifying and minimising the impact of fake visual media: Current and future directions. *Memory, Mind & Media, 1*, e15. [Crossref]

[69] Khoo, B., Phan, R. C. W., & Lim, C. H. (2022). Deepfake attribution: On the source identification of artificially generated images. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 12*(3), e1438. [Crossref]

[70] Masood, M., Nawaz, M., Malik, K. M., Javed, A., Irtaza, A., & Malik, H. (2023). Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward. *Applied intelligence, 53*(4), 3974-4026. [Crossref]

[71] Gupta, G., Raja, K., Gupta, M., Jan, T., Whiteside, S. T., & Prasad, M. (2023). A comprehensive review of deepfake detection using advanced machine learning and fusion methods. *Electronics, 13*(1), 95. [Crossref]

[72] Khan, S. A., & Dang-Nguyen, D. T. (2022, September). Hybrid transformer network for deepfake detection. In *Proceedings of the 19th international conference on content-based multimedia indexing* (pp. 8-14). [Crossref]

[73] Sharma, V. K., Garg, R., & Caudron, Q. (2024). A systematic literature review on deepfake detection techniques. *Multimedia Tools and Applications*, 1-43. [Crossref]

[74] Kharbat, F. F., Elamsy, T., Mahmoud, A., & Abdullah, R. (2019, November). Image feature detectors for deepfake video detection. In *2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA)* (pp. 1-4). IEEE. [Crossref]

[75] Kim, J., Kong, J., & Son, J. (2021, July). Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International Conference on Machine Learning* (pp. 5530-5540). PMLR.

[76] Kim, J., Lee, J., & Zhang, B. T. (2022, June). Smooth-Swap: A Simple Enhancement for Face-Swapping with Smoothness. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 10769-10778). IEEE. [Crossref]

[77] Kim, K. W., Park, S. W., Lee, J., & Joe, M. C. (2022, May). Assem-vc: Realistic voice conversion by assembling modern speech synthesis techniques. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6997-7001). IEEE. [Crossref]

[78] Kim, M., Liu, F., Jain, A., & Liu, X. (2023). Dcface: Synthetic face generation with dual condition diffusion model. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 12715-12725). [Crossref]

[79] Kim, M., Tariq, S., & Woo, S. S. (2021). Fretal: Generalizing deepfake detection using knowledge distillation and representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1001-1012). [Crossref]

[80] Kingma, D. P., & Welling, M. (2013, December). *Auto-encoding variational bayes*.

[81] Koizumi, Y., Zen, H., Karita, S., Ding, Y., Yatabe, K., Morioka, N., ... & Bapna, A. (2023). Libritts-r: A restored multi-speaker text-to-speech corpus. *arXiv preprint arXiv:2305.18802*. [Crossref]

[82] Koopman, M., Rodriguez, A. M., & Geradts, Z. (2018, August). Detection of deepfake video manipulation. In *The 20th Irish machine vision and image processing conference (IMVIP)* (pp. 133-136).

[83] Korshunov, P., & Marcel, S. (2018, September).

Speaker inconsistency detection in tampered video. In *2018 26th European signal processing conference (EUSIPCO)* (pp. 2375-2379). IEEE. [Crossref]

[84] Kumar, D. A., & Priyanka, K. (2025, January). Enhancing Emotional Voice Conversion with Intensity Control and Mixed Embedding. In *2025 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)* (pp. 1-8). IEEE. [Crossref]

[85] Kumar, P., Vatsa, M., & Singh, R. (2020). Detecting face2face facial reenactment in videos. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 2589-2597). [Crossref]

[86] Kurematsu, A., Takeda, K., Sagisaka, Y., Katagiri, S., Kuwabara, H., & Shikano, K. (1990). ATR Japanese speech database as a tool of speech recognition and synthesis. *Speech communication, 9*(4), 357-363. [Crossref]

[87] Lee, C. H., Liu, Z., Wu, L., & Luo, P. (2020). Maskgan: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5549-5558). [Crossref]

[88] Li, S., Liu, L., Liu, J., Song, W., Hao, A., & Qin, H. (2023). SC-GAN: Subspace clustering based GAN for automatic expression manipulation. *Pattern Recognition, 134*, 109072. [Crossref]

[89] Li, X., Li, N., Weng, C., Liu, X., Su, D., Yu, D., & Meng, H. (2021, June). Replay and synthetic speech detection with res2net architecture. In *ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 6354-6358). IEEE. [Crossref]

[90] Li, Y., Chang, M. C., & Lyu, S. (2018, December). In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In *2018 IEEE International workshop on information forensics and security (WIFS)* (pp. 1-7). IEEE. [Crossref]

[91] Li, Y., & Lyu, S. (2018). Exposing deepfake videos by detecting face warping artifacts. *arXiv preprint arXiv:1811.00656*. [Crossref]

[92] Liang, P., Liu, G., Xiong, Z., Fan, H., Zhu, H., & Zhang, X. (2023). A facial geometry based detection model for face manipulation using CNN-LSTM architecture. *Information Sciences, 633*, 370-383. [Crossref]

[93] Liu, R., Zhang, J., Gao, G., & Li, H. (2023). Betray oneself: A novel audio deepfake detection model via mono-to-stereo conversion. *arXiv preprint arXiv:2305.16353*. [Crossref]

[94] Liu, Z., Luo, P., Wang, X., & Tang, X. (2018). Large-scale celebfaces attributes (celeba) dataset. *Retrieved August, 15*(2018), 11.

[95] Lyth, D., & King, S. (2024). Natural language guidance of high-fidelity text-to-speech with synthetic annotations. *arXiv preprint arXiv:2402.01912*. [Crossref]

[96] Ma, H., Yi, J., Tao, J., Bai, Y., Tian, Z., & Wang, C. (2021). Continual learning for fake audio detection. *arXiv preprint arXiv:2104.07286*. [Crossref]

[97] Ma, X., Zhang, R., Wei, J., Lu, X., Xu, J., Zhang, L., & Lu, W. (2025). Self-distillation-based domain exploration for source speaker verification under spoofed speech from unknown voice conversion. *Speech Communication, 167*, 103153. [Crossref]

[98] Mandelli, S., Bonettini, N., Bestagini, P., & Tubaro, S. (2022, October). Detecting gan-generated images by orthogonal training of multiple cnns. In *2022 IEEE International Conference on Image Processing (ICIP)* (pp. 3091-3095). IEEE. [Crossref]

[99] Matern, F., Riess, C., & Stamminger, M. (2019, January). Exploiting visual artifacts to expose deepfakes and face manipulations. In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)* (pp. 83-92). IEEE. [Crossref]

[100] Mazaheri, G., & Roy-Chowdhury, A. K. (2022). Detection and localization of facial expression manipulations. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision* (pp. 1035-1045). [Crossref]

[101] McCloskey, S., & Albright, M. (2018). Detecting gan-generated imagery using color cues. *arXiv preprint arXiv:1812.08247*. [Crossref]

[102] Messer, K., Matas, J., Kittler, J., Luettin, J., & Maitre, G. (1999, March). XM2VTSDB: The extended M2VTS database. In *Second international conference on audio and video-based biometric person authentication* (Vol. 964, pp. 965-966).

[103] Mollahosseini, A., Hasani, B., & Mahoor, M. H. (2017). Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing, 10*(1), 18-31. [Crossref]

[104] Moschoglou, S., Papaioannou, A., Sagonas, C., Deng, J., Kotsia, I., & Zafeiriou, S. (2017). Agedb: the first manually collected, in-the-wild age database. In *proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 51-59). [Crossref]

[105] Müller, N. M., Czempin, P., Dieckmann, F., Froghyar, A., & Böttinger, K. (2022). Does audio deepfake detection generalize?. *arXiv preprint arXiv:2203.16263*. [Crossref]

[106] Nagahara, S., Katayama, T., Song, T., & Shimamoto, T. (2022, July). A Novel Video Coding Framework with GAN-based Face Generation for Videoconferencing. In *2022 37th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC)* (pp. 450-452). IEEE. [Crossref]

[107] Nagrani, A., Chung, J. S., & Zisserman, A. (2017). Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*. [Crossref]

[108] Nataraj, L., Mohammed, T. M., Chandrasekaran, S., Flenner, A., Bappy, J. H., Roy-Chowdhury, A. K., &

Manjunath, B. S. (2019). Detecting GAN generated fake images using co-occurrence matrices. *arXiv preprint arXiv:1903.06836*. [Crossref]

[109] Nawaz, M., Javed, A., & Irtaza, A. (2024). A deep learning model for FaceSwap and face-reenactment deepfakes detection. *Applied Soft Computing, 162*, 111854. [Crossref]

[110] Nguyen, H. H., Fang, F., Yamagishi, J., & Echizen, I. (2019, September). Multi-task learning for detecting and segmenting manipulated facial images and videos. In *2019 IEEE 10th international conference on biometrics theory, applications and systems (BTAS)* (pp. 1-8). IEEE. [Crossref]

[111] Kwak, J. G., Han, D. K., & Ko, H. (2020). CAFE-GAN: Arbitrary face attribute editing with complementary attention feature. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16* (pp. 524-540). Springer International Publishing. [Crossref]

[112] Oord, A. V. D., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., ... & Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*. [Crossref]

[113] Pal, M., Paul, D., & Saha, G. (2018). Synthetic speech detection using fundamental frequency variation and spectral features. *Computer Speech & Language, 48*, 31-50. [Crossref]

[114] Pal, M., Raikar, A., Panda, A., & Kopparapu, S. K. (2022). Synthetic speech detection using meta-learning with prototypical loss. *arXiv preprint arXiv:2201.09470*. [Crossref]

[115] Park, S. J., Kim, M., Hong, J., Choi, J., & Ro, Y. M. (2022, June). Synctalkface: Talking face generation with precise lip-syncing via audio-lip memory. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 36, No. 2, pp. 2062-2070). [Crossref]

[116] Parkhi, O., Vedaldi, A., & Zisserman, A. (2015). Deep face recognition. In *BMVC 2015-Proceedings of the British Machine Vision Conference 2015*. British Machine Vision Association.

[117] Pernuš, M., Štruc, V., & Dobrišek, S. (2023). MaskFaceGAN: High-resolution face editing with masked GAN latent code optimization. *IEEE Transactions on Image Processing, 32*, 5893-5908. [Crossref]

[118] Phillips, P. J., Flynn, P. J., Scruggs, T., Bowyer, K. W., Chang, J., Hoffman, K., ... & Worek, W. (2005, June). Overview of the face recognition grand challenge. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)* (Vol. 1, pp. 947-954). IEEE. [Crossref]

[119] Popov, V., Vovk, I., Gogoryan, V., Sadekova, T., & Kudinov, M. (2021, July). Grad-tts: A diffusion probabilistic model for text-to-speech. In *International conference on machine learning* (pp. 8599-8608). PMLR.

[120] Prajwal, K. R., Mukhopadhyay, R., Namboodiri, V. P., & Jawahar, C. V. (2020, October). A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM international conference on multimedia* (pp. 484-492). [Crossref]

[121] Pratap, V., Xu, Q., Sriram, A., Synnaeve, G., & Collobert, R. (2020). Mls: A large-scale multilingual dataset for speech research. *arXiv preprint arXiv:2012.03411*. [Crossref]

[122] Rana, M. S., & Sung, A. H. (2020, August). Deepfakestack: A deep ensemble-based learning technique for deepfake detection. In *2020 7th IEEE international conference on cyber security and cloud computing (CSCloud)/2020 6th IEEE international conference on edge computing and scalable cloud (EdgeCom)* (pp. 70-75). IEEE. [Crossref]

[123] Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 1-11). [Crossref]

[124] Saito, Y., Takamichi, S., Iimori, E., Tachibana, K., & Saruwatari, H. (2023). Chatgpt-edss: Empathetic dialogue speech synthesis trained from chatgpt-derived context word embeddings. *arXiv preprint arXiv:2305.13724*. [Crossref]

[125] Salvi, D., Bestagini, P., & Tubaro, S. (2023, June). Synthetic speech detection through audio folding. In *Proceedings of the 2nd ACM International Workshop on Multimedia AI against Disinformation* (pp. 3-9). [Crossref]

[126] Sanabria, R., Bogoychev, N., Markl, N., Carmantini, A., Klejch, O., & Bell, P. (2023, June). The edinburgh international accents of english corpus: Towards the democratization of english asr. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1-5). IEEE. [Crossref]

[127] Sarfjoo, S. S., Wang, X., Henter, G. E., Lorenzo-Trueba, J., Takaki, S., & Yamagishi, J. (2019). Transformation of low-quality device-recorded speech to high-quality speech using improved SEGAN model. *arXiv preprint arXiv:1911.03952*. [Crossref]

[128] Scherhag, U., Debiasi, L., Rathgeb, C., Busch, C., & Uhl, A. (2019). Detection of face morphing attacks based on PRNU analysis. *IEEE Transactions on Biometrics, Behavior, and Identity Science, 1*(4), 302-317. [Crossref]

[129] Sengupta, S., Chen, J. C., Castillo, C., Patel, V. M., Chellappa, R., & Jacobs, D. W. (2016, March). Frontal to profile face verification in the wild. In *2016 IEEE winter conference on applications of computer vision (WACV)* (pp. 1-9). IEEE. [Crossref]

[130] Shahzad, S. A., Hashmi, A., Khan, S., Peng, Y. T., Tsao, Y., & Wang, H. M. (2022, November). Lip sync matters: A novel multimodal forgery detector. In *2022 Asia-Pacific Signal and Information Processing Association*

Annual Summit and Conference (APSIPA ASC) (pp. 1885-1892). IEEE. [Crossref]

[131] Shao, M., Lu, L., Ding, Y., & Liao, Q. (2023, June). Minimising Distortion for GAN-Based Facial Attribute Manipulation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1-5). IEEE. [Crossref]

[132] Sisman, B., Yamagishi, J., King, S., & Li, H. (2020). An overview of voice conversion and its challenges: From statistical modeling to deep learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29*, 132-157. [Crossref]

[133] Solanki, G. K., & Roussos, A. (2022, October). Deep semantic manipulation of facial videos. In *European conference on computer vision* (pp. 104-120). Cham: Springer Nature Switzerland. [Crossref]

[134] Chung, J. S., Nagrani, A., & Zisserman, A. (2018). Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*. [Crossref]

[135] Sun, P., Qi, H., Li, Y., & Lyu, S. (2024). FakeTracer: Catching Face-swap DeepFakes via Implanting Traces in Training. *IEEE Transactions on Emerging Topics in Computing*, 13(1), 134-146. [Crossref]

[136] Suwajanakorn, S., Seitz, S. M., & Kemelmacher-Shlizerman, I. (2017). Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (ToG), 36*(4), 1-13. [Crossref]

[137] Tan, X., Chen, J., Liu, H., Cong, J., Zhang, C., Liu, Y., ... & Liu, T. Y. (2024). Naturalspeech: End-to-end text-to-speech synthesis with human-level quality. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 46*(6), 4234-4245. [Crossref]

[138] Tan, X., Qin, T., Soong, F., & Liu, T. Y. (2021). A survey on neural speech synthesis. *arXiv preprint arXiv:2106.15561*. [Crossref]

[139] Tanaka, K., Kameoka, H., & Kaneko, T. (2023). PRVAE-VC: Non-Parallel Many-to-Many Voice Conversion with Perturbation-Resistant Variational Autoencoder. In *12th Speech Synthesis Workshop (SSW) 2023*.

[140] Tang, J., Wang, K., Zhou, H., Chen, X., He, D., Hu, T., ... & Wang, J. (2025). Real-time neural radiance talking portrait synthesis via audio-spatial decomposition. *International Journal of Computer Vision*, 1-12. [Crossref]

[141] Thies, J., Zollhöfer, M., & Nießner, M. (2019). Deferred neural rendering: Image synthesis using neural textures. *Acm Transactions on Graphics (TOG), 38*(4), 1-12. [Crossref]

[142] Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., & Nießner, M. (2016). Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2387-2395). [Crossref]

[143] Todisco, M., Wang, X., Vestman, V., Sahidullah, M., Delgado, H., Nautsch, A., ... & Lee, K. A. (2019). ASVspoof 2019: Future horizons in spoofed and fake audio detection. *arXiv preprint arXiv:1904.05441*. [Crossref]

[144] Vecino, B. T., Pomirski, A., Iddon, T., Cotescu, M., & Lorenzo-Trueba, J. (2025). Lightweight End-to-end Text-to-speech Synthesis for low resource on-device applications. *arXiv preprint arXiv:2505.07701*. [Crossref]

[145] Vieira, T. F., Bottino, A., Laurentini, A., & De Simone, M. (2014). Detecting siblings in image pairs. *The Visual Computer, 30*, 1333-1345. [Crossref]

[146] Wang, C., Riviere, M., Lee, A., Wu, A., Talnikar, C., Haziza, D., ... & Dupoux, E. (2021). VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. *arXiv preprint arXiv:2101.00390*. [Crossref]

[147] Wang, F., Xiang, S., Liu, T., & Fu, Y. (2021, July). Attention based facial expression manipulation. In *2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)* (pp. 1-6). IEEE. [Crossref]

[148] Wang, J., Tondi, B., & Barni, M. (2022). An eyes-based siamese neural network for the detection of gan-generated face images. *Frontiers in Signal Processing, 2*, 918725. [Crossref]

[149] Wang, J., Tondi, B., & Barni, M. (2023, June). Classification of synthetic facial attributes by means of hybrid classification/localization patch-based analysis. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1-5). IEEE. [Crossref]

[150] Wang, K., Dunn, E., Rodriguez, M., & Frahm, J. M. (2017). Efficient video collection association using geometry-aware Bag-of-Iconics representations. *IPSJ Transactions on Computer Vision and Applications, 9*, 1-17. [Crossref]

[151] Wang, R., Juefei-Xu, F., Huang, Y., Guo, Q., Xie, X., Ma, L., & Liu, Y. (2020, October). Deepsonar: Towards effective and robust detection of ai-synthesized fake voices. In *Proceedings of the 28th ACM international conference on multimedia* (pp. 1207-1216). [Crossref]

[152] Wang, R., Juefei-Xu, F., Ma, L., Xie, X., Huang, Y., Wang, J., & Liu, Y. (2019). Fakespotter: A simple yet robust baseline for spotting ai-synthesized fake faces. *arXiv preprint arXiv:1909.06122*. [Crossref]

[153] Wang, T., Cheng, H., Zhang, X., & Wang, Y. (2025). NullSwap: Proactive Identity Cloaking Against Deepfake Face Swapping. *arXiv preprint arXiv:2503.18678*. [Crossref]

[154] Wang, Y., Skerry-Ryan, R. J., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., ... & Saurous, R. A. (2017). Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*. [Crossref]

[155] Waseem, S., Abu-Bakar, S. A. R. S., Omar, Z., Ahmed, B. A., Baloch, S., & Hafeezallah, A. (2023). Multi-attention-based approach for deepfake face and

expression swap detection and localization. *EURASIP Journal on Image and Video Processing, 2023*(1), 14. [Crossref]

[156] Wodajo, D., & Atnafu, S. (2021). Deepfake video detection using convolutional vision transformer. *arXiv preprint arXiv:2102.11126*. [Crossref]

[157] Wu, H., Kuo, H. C., Zheng, N., Hung, K. H., Lee, H. Y., Tsao, Y., ... & Meng, H. (2022, May). Partially fake audio detection by self-attention-based fake span discovery. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 9236-9240). IEEE. [Crossref]

[158] Wu, W., Qian, C., Yang, S., Wang, Q., Cai, Y., & Zhou, Q. (2018). Look at boundary: A boundary-aware face alignment algorithm. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2129-2138). [Crossref]

[159] Wu, X., Hu, P., Wu, Y., Lyu, X., Cao, Y. P., Shan, Y., ... & Qi, X. (2023). Speech2lip: High-fidelity speech to lip generation by learning from a short video. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 22168-22177). [Crossref]

[160] Wu, Z., Das, R. K., Yang, J., & Li, H. (2020). Light convolutional neural network with feature genuinization for detection of synthetic speech attacks. *arXiv preprint arXiv:2009.09637*. [Crossref]

[161] Wu, Z., Kinnunen, T., Evans, N., Yamagishi, J., Hanilçi, C., Sahidullah, M., & Sizov, A. (2015, September). ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge. In *INTERSPEECH 2015, Automatic Speaker Verification Spoofing and Countermeasures Challenge, colocated with INTERSPEECH 2015* (pp. 2037-2041). ISCA. [Crossref]

[162] Xiao, L., & Wang, Z. (2018, August). Dense convolutional recurrent neural network for generalized speech animation. In *2018 24th International Conference on Pattern Recognition (ICPR)* (pp. 633-638). IEEE. [Crossref]

[163] Xu, Z., Hong, Z., Ding, C., Zhu, Z., Han, J., Liu, J., & Ding, E. (2022, June). Mobilefaceswap: A lightweight framework for video face swapping. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 36, No. 3, pp. 2973-2981). [Crossref]

[164] Xu, Z., Zhou, H., Hong, Z., Liu, Z., Liu, J., Guo, Z., ... & Wang, J. (2022, October). Styleswap: Style-based generator empowers robust face swapping. In *European Conference on Computer Vision* (pp. 661-677). Cham: Springer Nature Switzerland. [Crossref]

[165] Xue, Z., Jiang, X., Liu, Q., & Wei, Z. (2023). Global–local facial fusion based GAN generated fake face detection. *Sensors, 23*(2), 616. [Crossref]

[166] Yadav, A. K. S., Xiang, Z., Bartusiak, E. R., Bestagini, P., Tubaro, S., & Delp, E. J. (2023, June). ASSD: Synthetic Speech Detection in the AAC Compressed Domain. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1-5). IEEE. [Crossref]

[167] Yamagishi, J., Wang, X., Todisco, M., Sahidullah, M., Patino, J., Nautsch, A., ... & Delgado, H. (2021). ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection. *arXiv preprint arXiv:2109.00537*. [Crossref]

[168] Yang, J., Das, R. K., & Li, H. (2018, November). Extended constant-Q cepstral coefficients for detection of spoofing attacks. In *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)* (pp. 1024-1029). IEEE. [Crossref]

[169] Yang, J., Das, R. K., & Li, H. (2019). Significance of subband features for synthetic speech detection. *IEEE Transactions on Information Forensics and Security, 15*, 2160-2170. [Crossref]

[170] Yang, J., Zhou, Y., & Huang, H. (2023). Mel-s3r: Combining mel-spectrogram and self-supervised speech representation with vq-vae for any-to-any voice conversion. *Speech Communication, 151*, 52-63. [Crossref]

[171] Yang, W., Zhou, X., Chen, Z., Guo, B., Ba, Z., Xia, Z., ... & Ren, K. (2023). Avoid-df: Audio-visual joint learning for detecting deepfake. *IEEE Transactions on Information Forensics and Security, 18*, 2015-2029. [Crossref]

[172] Yang, X., Li, Y., Qi, H., & Lyu, S. (2019, July). Exposing GAN-synthesized faces using landmark locations. In *Proceedings of the ACM workshop on information hiding and multimedia security* (pp. 113-118). [Crossref]

[173] Ye, Z., Jiang, Z., Ren, Y., Liu, J., He, J., & Zhao, Z. (2023). Geneface: Generalized and high-fidelity audio-driven 3d talking face synthesis. *arXiv preprint arXiv:2301.13430*. [Crossref]

[174] Yi, J., Bai, Y., Tao, J., Ma, H., Tian, Z., Wang, C., ... & Fu, R. (2021). Half-truth: A partially fake audio detection dataset. *arXiv preprint arXiv:2104.03617*. [Crossref]

[175] Yi, J., Fu, R., Tao, J., Nie, S., Ma, H., Wang, C., ... & Li, H. (2022, May). Add 2022: the first audio deep synthesis detection challenge. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 9216-9220). IEEE. [Crossref]

[176] Yi, J., Tao, J., Fu, R., Yan, X., Wang, C., Wang, T., ... & Li, H. (2023). Add 2023: the second audio deepfake detection challenge. *arXiv preprint arXiv:2305.13774*. [Crossref]

[177] Yin, F., Zhang, Y., Cun, X., Cao, M., Fan, Y., Wang, X., ... & Yang, Y. (2022, October). Styleheat: One-shot high-resolution editable talking face generation via pre-trained stylegan. In *European conference on computer vision* (pp. 85-101). Cham: Springer Nature Switzerland. [Crossref]

[178] Yoo, D., Lee, H., & Kim, J. (2024). Inversion based Face Swapping with Diffusion Model. *IEEE Access*, 13, 6764-6774. [Crossref]

[179] Yoo, S. M., Choi, T. M., Choi, J. W., & Kim, J. H. (2023). FastSwap: A lightweight one-stage framework for real-time face swapping. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 3558-3567). [Crossref]

[180] Yu, Y., Ni, R., Li, W., & Zhao, Y. (2022). Detection of AI-manipulated fake faces via mining generalized features. *ACM Transactions on Multimedia Computing, Communications*, and Applications (TOMM), 18(4), 1-23. [Crossref]

[181] Zen, H., Dang, V., Clark, R., Zhang, Y., Weiss, R. J., Jia, Y., ... & Wu, Y. (2019). Libritts: A corpus derived from librispeech for text-to-speech. *arXiv preprint arXiv:1904.02882*. [Crossref]

[182] Zhang, M., Zhou, Y., Ren, Y., Zhang, C., Yin, X., & Li, H. (2024). Refxvc: Cross-lingual voice conversion with enhanced reference leveraging. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32, 4146-4156. [Crossref]

[183] Li, S., Wu, S., Xiang, S., Zhang, Y., Guerrero, J. M., & Vasquez, J. C. (2020). Research on synchronverter-based regenerative braking energy feedback system of urban rail transit. *Energies, 13*(17), 4418. [Crossref]

[184] Zhang, Y., Zheng, L., & Thing, V. L. (2017, August). Automated face swapping and its detection. In *2017 IEEE 2nd international conference on signal and image processing* (*ICSIP*) (pp. 15-19). IEEE. [Crossref]

[185] Zhang, Z., Li, L., Ding, Y., & Fan, C. (2021). Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3661-3670). [Crossref]

[186] Zhao, Y., Huang, W. C., Tian, X., Yamagishi, J., Das, R. K., Kinnunen, T., ... & Toda, T. (2020). Voice conversion challenge 2020: Intra-lingual semi-parallel and cross-lingual voice conversion. *arXiv preprint arXiv:2008.12527*. [Crossref]

[187] Zheng, R., Song, B., & Ji, C. (2021, June). Learning pose-adaptive lip sync with cascaded temporal convolutional network. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*) (pp. 4255-4259). IEEE. [Crossref]

[188] Zheng, T., & Deng, W. (2018). Cross-pose lfw: A database for studying cross-pose face recognition in unconstrained environments. *Beijing University of Posts and Telecommunications, Tech. Rep, 5*(7), 5.

[189] Zheng, T., Deng, W., & Hu, J. (2017). Cross-age lfw: A database for studying cross-age face recognition in unconstrained environments. *arXiv preprint arXiv:1708.08197*. [Crossref]

[190] Zhong, W., Li, J., Cai, Y., Lin, L., & Li, G. (2024). Style-preserving lip sync via audio-aware style reference. *arXiv preprint arXiv:2408.05412*. [Crossref]

[191] Zhou, Q., Zhou, Z., Bao, Z., Niu, W., & Liu, Y. (2024). IIN-FFD: Intra-Inter Network for Face Forgery Detection. *Tsinghua Science and Technology, 29*(6), 1839-1850. [Crossref]

[192] Zhou, Y., Han, X., Shechtman, E., Echevarria, J., Kalogerakis, E., & Li, D. (2020). Makelttalk: speaker-aware talking-head animation. *ACM Transactions On Graphics (TOG), 39*(6), 1-15. [Crossref]

[193] Zhou, Y., & Lim, S. N. (2021). Joint audio-visual deepfake detection. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 14800-14809). [Crossref]

[194] Zhu, Y., Zhao, W., Tang, Y., Rao, Y., Zhou, J., & Lu, J. (2024). Stableswap: stable face swapping in a shared and controllable latent space. *IEEE Transactions on Multimedia*, 26, 7594-7607. [Crossref]

Arshad Ahmad is currently working as an Associate Professor at the Department of Computer Software Engineering, National University of Sciences & Technology, Pakistan. He holds a post-doctorate from the Institute of Software Systems Engineering, Johannes Kepler University, Linz, Austria. He received the MS Software Engineering degree from Blekinge Tekniska Högskola (BTH), Sweden in 2008. He worked as a Research Assistant at Fraunhofer Institute of Experimental Software Engineering (IESE), Germany and Vienna University of Technology, Austria respectively during the years 2010-2014. Afterwards, he received the PhD degree in Computer Science & Technology (specialization in Software Engineering) from Beijing Institute of Technology, China in 2018. He worked as an Assistant Professor of Software Engineering & Computer Science at Sino-Pak Center for Artificial Intelligence and Department of IT & Computer Science, Pak Austria Fachhochschule: Institute of Applied Sciences & Technology, Haripur, Pakistan from October 2020. Previously, he served as an Assistant Professor of Computer Science at University of Swabi, Swabi and City University of Science & Information Technology, Peshawar, respectively during the years 2019-2020. He has published several research papers in well reputed peer reviewed international journals and conferences. His current research interests include requirements engineering, text mining, opinion mining, sentiment analysis and machine learning, among others.