forensic sciences

MDPI

Review

Video and Audio Deepfake Datasets and Open Issues in Deepfake Technology: Being Ahead of the Curve

Zahid Akhtar [1,*], Thanvi Lahari Pendyala [2] and Virinchi Sai Athmakuri [1]

[1] Department of Network and Computer Security, State University of New York (SUNY) Polytechnic Institute, Utica, NY 13502, USA; athmakv@sunypoly.edu
[2] Department of Computer Science, State University of New York (SUNY) Polytechnic Institute, Utica, NY 13502, USA; pendyat@sunypoly.edu
* Correspondence: akhtarz@sunypoly.edu; Tel.: +1-315-792-7238

Abstract: The revolutionary breakthroughs in Machine Learning (ML) and Artificial Intelligence (AI) are extensively being harnessed across a diverse range of domains, e.g., forensic science, healthcare, virtual assistants, cybersecurity, and robotics. On the flip side, they can also be exploited for negative purposes, like producing authentic-looking fake news that propagates misinformation and diminishes public trust. Deepfakes pertain to audio or visual multimedia contents that have been artificially synthesized or digitally modified through the application of deep neural networks. Deepfakes can be employed for benign purposes (e.g., refinement of face pictures for optimal magazine cover quality) or malicious intentions (e.g., superimposing faces onto explicit image/video to harm individuals producing fake audio recordings of public figures making inflammatory statements to damage their reputation). With mobile devices and user-friendly audio and visual editing tools at hand, even non-experts can effortlessly craft intricate deepfakes and digitally altered audio and facial features. This presents challenges to contemporary computer forensic tools and human examiners, including common individuals and digital forensic investigators. There is a perpetual battle between attackers armed with deepfake generators and defenders utilizing deepfake detectors. This paper first comprehensively reviews existing image, video, and audio deepfake databases with the aim of propelling next-generation deepfake detectors for enhanced accuracy, generalization, robustness, and explainability. Then, the paper delves deeply into open challenges and potential avenues for research in the audio and video deepfake generation and mitigation field. The aspiration for this article is to complement prior studies and assist newcomers, researchers, engineers, and practitioners in gaining a deeper understanding and in the development of innovative deepfake technologies.

Keywords: deepfakes; digital face manipulations; computer forensics; digital forensics; deepfake generation; deepfake detection; fake news; multimedia manipulations; information authenticity

check for updates

Citation: Akhtar, Z.; Pendyala, T.L.; Athmakuri, V.S. Video and Audio Deepfake Datasets and Open Issues in Deepfake Technology: Being Ahead of the Curve. *Forensic Sci.* **2024**, *4*, 289–377. https://doi.org/10.3390/forensicsci4030021

Academic Editors: Mary Aiken and Ricardo Jorge Dinis-Oliveira

Received: 11 January 2024
Revised: 5 July 2024
Accepted: 5 July 2024
Published: 13 July 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The rapid advancements in artificial intelligence and machine learning are widely being exploited by a wide variety of technologies, including good (e.g., smart home) and bad (e.g., generating sophisticated cyberattacks) applications. One such harmful application involves deepfakes and digital manipulations of audio and facial visual contents. 'Deepfake' is coined by fusing 'deep learning' and 'fake'. It is estimated that a significant portion (e.g., around 50% [1]) of billions of audio, images, and videos that are uploaded daily to various online platforms, spanning social and professional networking sites, appears to be manipulated [2,3]. As faces and speeches are integral to human interactions and serve as the foundation for biometrics-based person recognition, manipulated faces and speeches pose remarkable troubles to Internet information integrity and security systems [4].

Deepfakes are believable audio-, visual-, or multimedia-content that are digitally modified or synthetically generated through the application of AI and deep learning models [5–7]. Digital face manipulation encompasses the alteration of facial traits, e.g., gender,

ethnicity, age, eyeglasses, emotion, morphing, beard, makeup, mustache, simulated effects of drug use, attractiveness, mouth closed or open, hair color/length/style, gaze, injury, skin texture or color, pose, adversarial examples (i.e., adding imperceptible perturbations), and eye color [2,8,9], as depicted in Figure 1. All in all, face manipulation or deepfakes could be grouped into four main categories: identity swap, face reenactment, attribute manipulation, and non-existing entire audio or face synthesis [5], as presented in Figure 2.

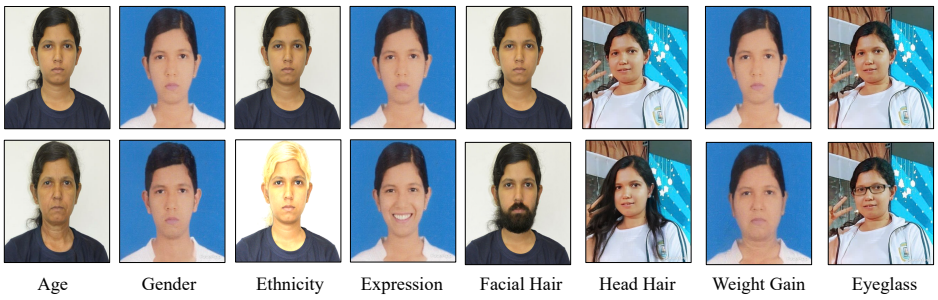


Figure 1. Illustrative instances of distinct face manipulations: the first row features original samples while the second row presents manipulated counterparts created using mobile apps like FaceApp [10].

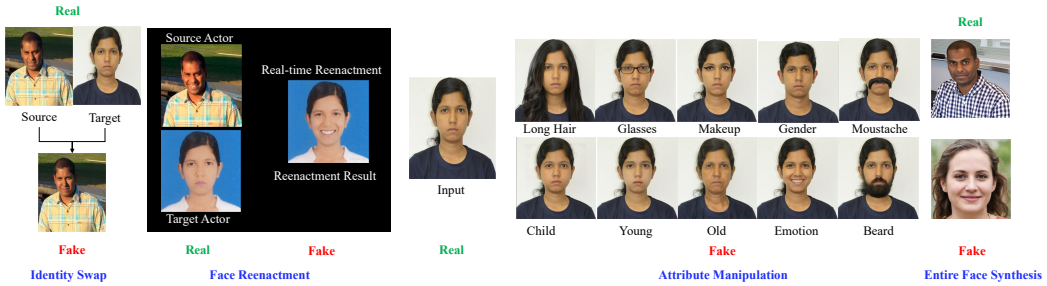


Figure 2. Authentic and fake illustrations from various deepfake or face manipulation categories. In the "Entire Face Synthesis" group, the fake example was derived from the approach outlined in [11].

Similarly, audio deepfakes involve manipulating or synthesizing speech samples in a way that changes the original sentences or generating utterly new audio contents of a particular individual [12,13]. As depicted in Figure 3, the main types of audio deepfakes [14–17] are voice cloning (i.e., generating synthetic voice recordings, which closely sound like a specific person, using deep learning (DL) models trained on the target's voice data), speech synthesis (i.e., generating natural-sounding speech from text input, also referred to as text-to-speech (TTS) synthesis), voice conversion (i.e., altering the speech attributes of a source speaker to make it sound like target person's voice while preserving linguistic original speech's content, also referred to as impersonation or imitation-based audio deepfake), audio manipulation (i.e., modifying speech by changing tone, pitch, tempo, emotional expression, rearranging, adding or removing words and sentences, adding or removing background noise, or converting accent or gender), language translation (i.e., converting speech from one language to a different language while preserving the original speaker's voice characteristics), and half deepfake (i.e., manipulating and synthesizing a specific segment of the audio while preserving the rest of the original audio source unchanged, also referred to as partial deepfake).

The prevalence of affordable and high-tech mobile devices (e.g., smartphones) along with easily accessible face and audio editing apps (e.g., FaceApp [10] and ResembleAI [18]) as well as deep neural network source codes (e.g., Identity-aware Dynamic Network [19] and vocoder-network-based multispeaker text-to-speech synthesis (SV2TTS) [20]) has empowered even non-experts to produce intricate deepfakes and digitally modified face samples that pose challenges for current computer forensic tools and human examiners.

The ongoing deepfake evolution is signifying the emergence of fake news 2.0 (e.g., more evolved fake news using highly realistic deepfakes created via sophisticated AI), disinformation and misinformation to be exploited by various entities such as bots, foreign governments, hyperpartisan media, conspiracy theorists, and trolls. For instance, a chief executive officer (CEO) was scammed into losing USD 243,000 via the use of an audio deepfake [21].

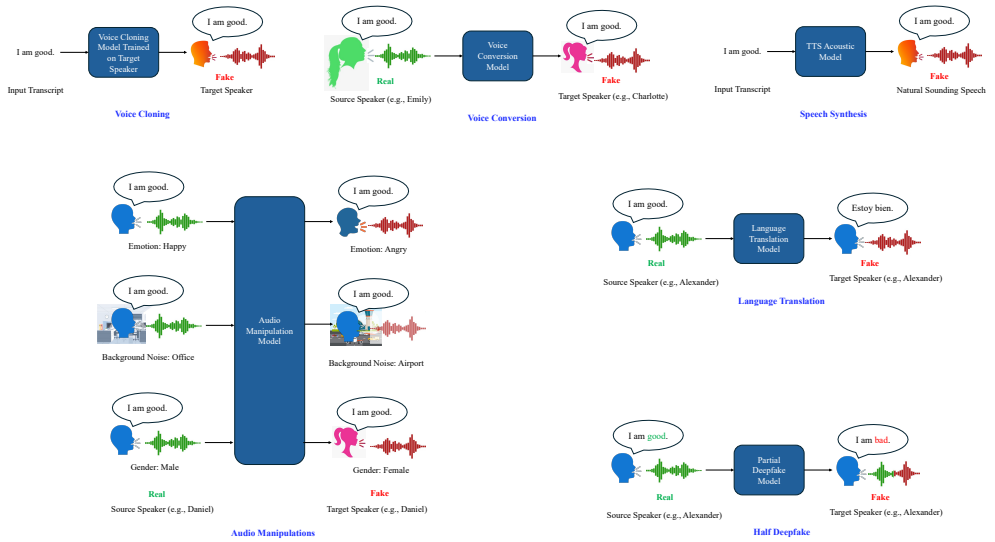


Figure 3. An illustration of audio deepfake categories.

In recent years, in the face deepfake field, several methods for identity swap (deepfake) generation (e.g., Shu et al. [22]), identity swap (deepfake) detection (e.g., Wang et al. [23]), reenactment generation (Agarwal et al. [24]), reenactment detection (e.g., Cozzolino et al. [25]), attribute manipulation generation (e.g., Patashnik et al. [26]), attribute manipulation detection (e.g., Asnani et al. [27]), entire face synthesis generation (e.g., Li et al. [28]), and entire face synthesis detection (e.g., Tan et al. [29]) have been developed. Similarly, in the audio deepfake field, several methods for voice cloning generation (e.g., Luong et al. [30]), voice cloning detection (e.g., Kulangareth et al. [31]), speech synthesis generation (e.g., Oord et al. [32]), speech synthesis detection (e.g., Rahman et al. [33]), voice conversion generation (e.g., Wang et al. [34]), voice conversion detection (e.g., Lo et al. [35]), audio manipulation generation (e.g., Choi et al. [36]), audio manipulation detection (e.g., Zhao et al. [37]), language translation generation (e.g., Jia et al. [38]), language translation detection (e.g., Kuo et al. [39]), half deepfake generation (e.g., Yi et al. [40]), and half deepfake detection (e.g., Wu et al. [41]) have been formulated. Despite the great progress, most audio and video deepfake detection frameworks have a lower generalization capability (i.e., their accuracy substantially decreases under novel deepfakes [42]), are not useful for real-time mobile deepfake detection [43], and are susceptible to adversarial attacks [44], primarily owing to their reactive nature rather than being proactive. It is a continuous arms race between attackers (i.e., deepfake generation techniques) and defenders (i.e., deepfake detection techniques). It is imperative for the researchers and professionals to stay abreast of the latest deepfake advancements in order to secure victory in this crucial race.

Consequently, a few survey articles have been produced in the literature (e.g., [5,45–48]). However, they did not present systematic and detailed information about different deepfake datasets as well as comprehensive open issues and challenges for academics and practitioners in the multidisciplinary fields. To bridge this gap, this paper provides the following: (i) a comprehensive overview of existing image, video, and audio deepfake databases that could be utilized not only for enhancing accuracy, generalization, and resilience against attacks of deepfake detection techniques but also for devising novel deepfake generation methods and tools; (ii) an extensive discussion of open challenges

and potential research directions in the field of audio and visual deepfake generation and mitigation. All in all, we hope that this paper will complement prior survey articles and help newcomers, researchers, and engineers achieve profound comprehension and novel deepfake algorithms developments.

The rest of the paper is organized as follows: Section 2 presents the overview of available image and video datasets employed for deepfakes mitigation techniques. Section 3 provides a review of available audio databases utilized for audio deepfakes detection methods. Section 4 is dedicated to discussing the open issues in deepfake technology and exploring possible future directions. Section 5 outlines the key conclusions.

2. Image and Video Deepfake Datasets

Deepfake databases play a pivotal role in training, testing, and benchmarking countermeasures against deepfakes. The presence of an array of diverse databases will help advancing deepfake technology. In this section, we present a comprehensive overview of exiting datasets focused on image and video deepfakes. We thoroughly examined various research articles and database repositories to furnish comprehensive details, including information not typically found in previous works, such as the memory requirements for downloading or saving databases. We have incorporated all publicly reported and/or accessible datasets from 2013 to 2024, illustrating the comprehensive advancements within the image and video deepfake datasets field. Tables 1–21 present comparative analysis of existing datasets for image and video deepfakes.

Tables 1–21 list the image and video deepfake datasets in ascending order by year (first column) and then in alphabetical order by name within each year (second column) to provide a view of how the datasets and the deepfake field have progressed over time. It is hoped that this chronological arrangement will help readers observe the evolution of image and video deepfake datasets, deepfake types, unimodal or multimodal characteristics, quantities, quality, and the methods/tools used to generate deepfakes.

2.1. DSI-1 [49]

This dataset comprises 25 authentic and 25 manipulated images sourced from diverse websites, each possessing different resolutions. The images were altered by adding one or more individuals and exhibit medium to high visual quality. A splicing technique was used to create the deepfakes.

2.2. DSO-1 [49]

This dataset is composed of 100 real and 100 fake images. The images were manipulated by addition of one or more persons. Furthermore, splicing with brightness and color adjustments were also performed as part of the forgeries. The images have a resolution of 2048×1536 pixels.

2.3. CelebA-HQ [50]

This dataset features a collection of 30,000 high-quality deepfake celebrity images in JPG format. Every image is characterized by a pixel resolution of 1024×1024. The GAN technique was used to generate these deepfakes. This is a high-quality version deepfake dataset generated using CelebA dataset as its source. It has a total size of $\sim$28.21 gigabytes.

2.4. DeepfakeTIMIT [51]

This database was created in 2018 with 320 real and 620 deepfake videos. The deepfakes were created based on face swapping using the FSGAN (Face Swapping Generative Adversarial Network [52]) technique. The deepfakes are of two visual qualities, i.e., 64×64 pixels (low resolution) and 128×128 pixels (high resolution). The videos are in AVI format. The deepfakes are mostly with frontal faces and blur. There is no audio manipulation in the deepfakes. It has a size of 226.6 MB.

Table 1. Comparison and key characteristics of existing face image and video deepfake datasets. I = Images; V = Videos; C = Value or Visual Quality estimated by authors.

Year	Dataset	No. of Real Samples	No. of Fake Samples	Total No. of Samples	Types of Deepfakes	Visual Quality	Format	Resolution	Methods Used to Generate Deepfake	Size of Dataset	Link to Download the Dataset
2013	DSI-1 [49]	25 (I) [49]	25 (I) [49]	50 (I) (C)	Adding one or more individuals [49]	Medium to High (C)	-	Different resolutions [49]	Splicing [49]	-	-
2013	DSO-1 [49]	100 (I) [49]	100 (I) [49]	200 (I) (C)	Adding one or more individuals [49]	High (C)	-	2048 × 1536 [49]	Splicing and postprocessing [49]	-	-
2018	CelebA-HQ [50]	-	30,000 (I) [50]	30,000 (I) [50]	Entire Face Synthesis [50]	High [50]	JPG [50]	1024 × 1024 [50]	GAN [50]	∼28.21 GB [53]	https://drive.google.com/drive/folders/0B4qLcYyJmiz0TXY1NG02bzZVRGs?resourcekey=0-arAVTUfW9KRhN-irJchVKQ (accessed on 1 January 2024)
2018	Deepfake-TIMIT [51]	320 (V) [51]	620 (V) [51]	940 (V) (C)	Face Swap [51]	Low and High [51]	AVI [54]	64 × 64 (low), 128 × 128 (high) [51]	FaceSwap-GAN [51]	226.6 MB [54]	https://zenodo.org/records/4068245 (accessed on 31 December 2023)

Table 2. Comparison and key characteristics of existing face image and video deepfake datasets. I = Images; V = Videos; C = Value or Visual Quality estimated by authors.

Year	Dataset	No. of Real Samples	No. of Fake Samples	Total No. of Samples	Types of Deepfakes	Visual Quality	Format	Resolution	Methods Used to Generate Deepfake	Size of Dataset	Link to Download the Dataset
2018	EBV [55]	50 (V) [55]	49 (V) [55]	99 (C)	Face Swapping [55]	Low (C)	-	224 × 224 [55]	DeepFake algorithm with postprocessing [55]	1.01 GB [56]	One drive: https://1drv.ms/u/s!As4tun0sWfKsgdVcYJ-nn0bw0kdjzw?e=DAcGfb (accessed on 1 January 2024)
2018	Faceforensics [57]	1004 (V) [57]	1004 (V) [58]	2008 (V) (C)	Self-Reenactment, Source-to-Target [57]	High [57]	-	128 × 128 [57]	Face2Face [57,59]	Lossless compressed videos: 130 GB, Raw videos: 3.5 TB [60]	https://kaldir.vc.in.tum.de/faceforensics_download_v1.py (accessed on 31 December 2023)
2018	FFW [61]	-	53,000 (I) from 150 (V) [61]	53,000 (I) from 150 (V) [61]	Face Replacement, Image tampering [61]	Low (C)	MP4, AVI [62]	>= 480 [61]	CGI, GANs, FakeApp mobile application [61]	-	https://github.com/AliKhoda/FFW/blob/main/all_videos.csv (accessed on 1 January 2024)
2018	HOHA-based [63]	300 (V) [63]	300 (V) [63]	600 (V) [63]	Face Swapping [63]	Low or Medium (C)	MP4	299 × 299 [63]	Encoder–Decoder [63]	-	-

Table 3. Comparison and key characteristics of existing face image and video deepfake datasets. I = Images; V = Videos; C = Value or Visual Quality estimated by authors.

Year	Dataset	No. of Real Samples	No. of Fake Samples	Total No. of Samples	Types of Deepfakes	Visual Quality	Format	Resolution	Methods Used to Generate Deepfake	Size of Dataset	Link to Download the Dataset
2018	UADFV [64]	49 (V) [64]	49 (V) [64]	98 (V) [65]	Face Swap [66]	Low [66]	MP4 [67]	294 × 500 [64]	FakeApp mobile application [66]	146 MB [67]	https://drive.google.com/file/d/17d-0K2UblFldBmjTUk3_nASK8MhhiSHa/view (accessed on 31 December 2023)
2019	Celeb-DF [68]	590 (V) [68]	5639 (V) [68]	6229 (V) (C)	Face swap [68]	High [66,68]	MPEG4.0 [68]	256 × 256 [68]	Improved deepfake synthesis algorithm [68]	9.3 GB [69]	https://drive.google.com/file/d/1iLx76wsbi9itnkxSqz9BVBl4ZvnbIazj/view (accessed on 31 December 2023)
2019	DeepFake-Detection [70] (Google and Jigsaw dataset)	363 (V) [58,71]	3068 (V) [58,71]	3431 (V) (C)	Face swap [70]	Low [66]	-	-	DeepFakes (DF), Face2Face (F2F), FaceSwap (FS), Neural-Textures (NT) [58]	Source videos: raw/0: 200 GB; C23: 3 GB; C40: 400 MB; manipulated videos: raw/0: 1.6 TB; c23: 22 GB; c40: 3 GB [71]	https://kaldir.vc.in.tum.de/faceforensics_download_v4.py (accessed on 1 January 2024)

Table 4. Comparison and key characteristics of existing face image and video deepfake datasets. I = Images; V = Videos; C = Value or Visual Quality estimated by authors.

Year	Dataset	No. of Real Samples	No. of Fake Samples	Total No. of Samples	Types of Deepfakes	Visual Quality	Format	Resolution	Methods Used to Generate Deepfake	Size of Dataset	Link to Download the Dataset
2019	DEFACTO [72]	-	229,000 (I) [72]	229,000 (I) [72]	Face swapping, Face morphing [72]	Low (C)	TIF, JPG [73]	240 × 320–640 × 640 [73]	Copy–Move, splicing, object-removal, morphing [72]	121 GB [73]	https://www.kaggle.com/defactodataset/datasets (accessed on 31 December 2023)
2019	DFFD [74]	1000 (V), 58,703 (I) [74]	3000 (V), 240,336 (I) [74]	4000 (V), 299,039 (I) [74]	Identity swap, Expression swap, Attribute manipulation, Entire face Synthesis [74]	Low and High (C)	PNG, MP4 [75]	299 × 299 (I) [75]	FaceSwap, Deepfake, DeepFaceLab, Face2Face, FaceAPP, StarGAN, PGGAN, StyleGAN [74]	18.73 GB [75]	https://cvlab.cse.msu.edu/dffd-dataset.html (accessed on 31 December 2023)
2019	FaceForensics++ [76]	1000 (V) [76]	4000 (V) [76]	5000 (V) (C)	Face reconstruction, Face swap, Facial reenactment [76]	Low and High [76]	MP4, PNG [60]	854 × 480 (VGA), 280 × 720 (HD), 920 × 1080 (Full HD) [77]	DeepFakes (DF), Face2Face (F2F), FaceSwap (FS), NeuralTextures (NT) [76]	Original videos: 38.5 GB; All h264 compressed videos with compression rate factor: raw/0: 500 GB, 23: 10 GB, 40: 2 GB; All raw extracted images as PNGs: 2TB [60]	https://kaldir.vc.in.tum.de/faceforensics_download_v4.py (accessed on 31 December 2023)

Table 5. Comparison and key characteristics of existing face image and video deepfake datasets. I = Images; V = Videos; C = Value or Visual Quality estimated by authors.

Year	Dataset	No. of Real Samples	No. of Fake Samples	Total No. of Samples	Types of Deepfakes	Visual Quality	Format	Resolution	Methods Used to Generate Deepfake	Size of Dataset	Link to Download the Dataset
2019	FFHQ [78]	70,000 (I) [78]	140,000 (I) [79]	210,000 (I) [79]	Entire face synthesis, Attribute manipulation [78]	High [78]	PNG [79]	1024 × 1024 [78]	StyleGAN [78]	2.56 TB [79]	https://drive.google.com/drive/folders/1u2xu7bSrWxrbUxk-dT-UvEJq8IjdmNTP (accessed on 1 January 2024)
2019	WhichFaceReal [80]	70,000 (I) [80]	70,000 (I) (C)	140,000 (I) (C)	Face Synthesis [80]	High (C)	JPG [80]	1024 × 1024 (C)	StyleGAN [80]	-	https://www.whichfaceisreal.com/ (accessed on 1 January 2024)
2020	Deeper-Forensics-1.0 [81,82]	48,475 (V) [83]	11,000 (V) [83]	59,475 (V) (C)	Face swap [81]	High [66]	MP4 [84]	1920 × 1080 [81]	DF-VAE (DeepFake Variational Auto Encoder) [81]	281.35 GB [84]	https://drive.google.com/drive/folders/1s3KwYyTIXT78VzkRazn9QDPuNh18TWe- (accessed on 31 December 2023)
2020	DFDC [85]	23,654 (V) (C)	104,500 (V) [85]	128,154 (V) [85]	Face reenactment, Face swap [85]	High [66,85]	MP4 [86]	128 × 128 (DF-128) and 256 × 256 (DF-256) [85]	DFAE, MM/NN face swap, NTH, FSGAN, StyleGAN, Refinement, Audio Swaps [85]	471.84 GB [86]	https://www.kaggle.com/competitions/deepfake-detection-challenge/data (accessed on 31 December 2023)

Table 6. Comparison and key characteristics of existing face image and video deepfake datasets. I = Images; V = Videos; C = Value or Visual Quality estimated by authors.

Year	Dataset	No. of Real Samples	No. of Fake Samples	Total No. of Samples	Types of Deepfakes	Visual Quality	Format	Resolution	Methods Used to Generate Deepfake	Size of Dataset	Link to Download the Dataset
2020	DMD [87]	751 (I) [87]	8857 (I) [87]	9608 (I) [87]	Face swapping, Gender conversion, Face morphing, Smile manipulation [87]	Low (C)	JPG	640 × 480	FaceApp mobile application [87]	1.42 GB	-
2020	FaceShifter [88]	1000 (V) [66]	10,000 (V) [66]	11,000 (V) (C)	Face swap [88]	High [66]	-	256 × 256 [88]	FaceShifter [88]	-	https://lingzhili.com/FaceShifterPage/ (accessed on 1 January 2024)
2020	FakeET [89]	331 (V) [89]	480 (V) [89]	811 (V) [89]	Face swap [89]	High (C)	MP4 [90]	1920 × 1080 [89]	Collected from Google or Jigsaw dataset [89]	25.9 GB [90]	https://drive.google.com/drive/folders/1DpDIbjRTn3rTVdc5PU9uprRdLfmRgr-8?usp=sharing_eil_m&ts=655e5535 (accessed on 1 January 2024)

Table 7. Comparison and key characteristics of existing face image and video deepfake datasets. I = Images; V = Videos; C = Value or Visual Quality estimated by authors.

Year	Dataset	No. of Real Samples	No. of Fake Samples	Total No. of Samples	Types of Deepfakes	Visual Quality	Format	Resolution	Methods Used to Generate Deepfake	Size of Dataset	Link to Download the Dataset
2020	FFIW 10K [91]	10,000 (V) [91]	10,000 (V) [91]	20,000 (C)	Face swap [91]	High [66]	MP4 [92,93]	480 and above [91]	DeepFaceLab, FS-GAN, FaceSwap [91]	Train: 17 GB [92], Test: 4.1 GB [93]	Train + Val: https://drive.google.com/file/d/1-Ha_A9yRFS0dACrv-L156Kfy_yaPn980/view?usp=sharing; Test: https://drive.google.com/file/d/1ydNrV_LK3Ep6i3_WPsUo0_aQan4kDUbQ/view?usp=sharing
2020	iFakeFaceDB [94]	-	87,000 (I) [95]	87,000 (I) [95]	Entire face synthesis [95]	High (C)	JPG [96]	224 × 224 [95]	StyleGAN, GAN-PrintR [95]	1.4 GB [96]	http://socia-lab.di.ubi.pt/~jcneves/iFakeFaceDB.zip (accessed on 1 January 2024)
2020	RFFD [97]	Train: 1081 (I) [97]	Train: 960 (I) [97]	Train: 2041 (I) (C) [97]	Attribute manipulation [97]	Medium to High (C)	JPG [97]	600 × 600 [97]	Photoshop [97]	431 MB [97]	https://www.kaggle.com/datasets/ciplab/real-and-fake-face-detection (accessed on 1 January 2024)

Table 8. Comparison and key characteristics of existing face image and video deepfake datasets. I = Images; V = Videos; C = Value or Visual Quality estimated by authors.

Year	Dataset	No. of Real Samples	No. of Fake Samples	Total No. of Samples	Types of Deepfakes	Visual Quality	Format	Resolution	Methods Used to Generate Deepfake	Size of Dataset	Link to Download the Dataset
2020	UIBVFED [98]	-	640 (I) [98]	640 (I) [98]	Attribute manipulation, Entire Face Synthesis [98]	Medium to High (C)	PNG [98]	750 × 133 [98]	Blendshapes, Autodesk Character generator [98]	652 MB [99]	https://ugivia.uib.es/uibvfed/ (accessed on 1 January 2024)
2020	WildDeepfake [100]	3805 (V) [100]	3509 (V) [100]	7314 (C)	Face swap (deepfake) [100]	High [66]	PNG [101]	224 × 224 [100]	Collected from sources on Internet [100]	67.8 GB [101]	https://drive.google.com/drive/folders/1Cb_OqksBU3x7HFIo8EvDTigU6IjM7tmp (accessed on 31 December 2023)
2020	YouTube-DF [102]	98(V) [102]	98 (face-swaps), 79 (V) [102]	98 (face-swaps); 177 (V) [102]	Face swapping [102]	Low and High [102]	MP4 (C)	256 × 256 [102]	DeepFaceLab [102]		https://cs.uef.fi/deepfake_dataset/ (accessed on 1 January 2024)
2021	DeepFake MNIST+ [103]	10,000 (V) [103]	10,000 (V) [103]	20,000 (V) [103]	Image animation [103]	High [66]	MP4 [104]	256 × 256 [103]	First-Order Motion Model (FOMM) [103,105]	2.21 GB [104]	https://github.com/huangjiadidi/DeepFakeMnist (accessed on 31 December 2023)

Table 9. Comparison and key characteristics of existing face image and video deepfake datasets. I = Images; V = Videos; C = Value or Visual Quality estimated by authors.

Year	Dataset	No. of Real Samples	No. of Fake Samples	Total No. of Samples	Types of Deepfakes	Visual Quality	Format	Resolution	Methods Used to Generate Deepfake	Size of Dataset	Link to Download the Dataset
2021	DeepStreets [106]	600 (V) [106]	600 (V) [106]	1200 (V) [106]	Video to video synthesis [106]	High/ Low [106]	MP4 [107]	512 × 1024 [106]	Vid2vid, Wcvid2vid [106]	24.1 GB [107]	http://clem.dii.unisi.it/~vipp/datasets.html (accessed on 1 January 2024)
2021	DFGC-21 [108]	1000 (I) [108]	N × 1000 (I) [108]	Fake: N × 1000; Real: 1000 (I) (C)	Face swapping [108]	High (C)	PNG [108]	300 × 300 [108]	FaceShifter, FaceController, FaceSwap, FirstOrderMotion, FSGAN [108]	5.18 GB [109]	https://drive.google.com/drive/folders/1SD4L3R0XCZnr-LnZy5G9Vsho9BpIYe6Z (accessed on 1 January 2024)
2021	DF-Mobio [110]	31,950 (V) [110]	14,546 (V) [110]	46,496 (V) (C)	Face swap [110]	High (C)	MP4 [111]	256 × 256 [110]	GAN [110]	170.1 GB [111]	https://zenodo.org/records/5769057 (accessed on 3 July 2024)
2021	DF-W [112]	-	1869 (V) [112]	1869 (V) [112]	Face swap [112]	High (C)	MP4 [113]	360 × 360 (low) to 2560 × 1080 (high) [112]	Collected from YouTube, Reddit, and Bilibili [112]	31.55 GB [113]	https://github.com/jmpu/webconf21-deepfakes-in-the-wild (accessed on 1 January 2024)

Table 10. Comparison and key characteristics of existing face image and video deepfake datasets. I = Images; V = Videos; C = Value or Visual Quality estimated by authors.

Year	Dataset	No. of Real Samples	No. of Fake Samples	Total No. of Samples	Types of Deepfakes	Visual Quality	Format	Resolution	Methods Used to Generate Deepfake	Size of Dataset	Link to Download the Dataset
2021	FaceSynthesis [114]	-	100,000 (I) [114]	100,000 (I) [114]	Face synthesis [114]	High (C)	PNG [115]	512 × 512 [114]	VFX (3D face model) [114]	31.8 GB [115]	https://facesyntheticspubwedata.blob.core.windows.net/iccv-2021/dataset_100000.zip (accessed on 1 January 2024)
2021	FakeAV-Celeb [116]	500 (V) [116]	19,500 (V) [116]	20,000(V) [116]	Face swapping, facial reenactment, voice cloning [116]	High [116]	MP4 [117]	224 × 224 [118]	Video: FSGAN, Wav2Lip, FaceSwap; Audio: SV2TTS [116]	6 GB [118]	https://sites.google.com/view/fakeavcelebdash-lab/download?authuser=0 (accessed on 1 January 2024)
2021	ForgeryNet [119]	1,438,201 (I), 99,630 (V) [119]	1,457,861 (I), 121,617 (V) [119]	2,896,062 (I) (C), 221,247 (V) (C)	ID replaced: Face swap, Face transfer, Face stacked manipulation; ID remained: Face reenactment, Face editing [119]	High/Low [66]	JPG, MP4 [120]	240 to 1080 [119]	ID replaced: FSGAN, FaceShifter, BlendFace, MM Replacement; ID remained: MASKGAN, StarGAN, StyleGAN, ATVG-Net, SC-FEGAN [119]	496.23 GB [120]	Link 1: https://opendatalab.com/OpenDataLab/ForgeryNet/tree/main; Link 2: https://115.com/s/swnk84d3wl3?password=cvpr&#ForgeryNet (accessed on 31 December 2023)

Table 11. Comparison and key characteristics of existing face image and video deepfake datasets. I = Images; V = Videos; C = Value or Visual Quality estimated by authors.

Year	Dataset	No. of Real Samples	No. of Fake Samples	Total No. of Samples	Types of Deepfakes	Visual Quality	Format	Resolution	Methods Used to Generate Deepfake	Size of Dataset	Link to Download the Dataset
2021	HiFiFace [121]	-	1000 (V) [122]	1000 (V) [122]	Face swapping [121]	High [121]	MP4 [122]	256 × 256 and 512 × 512 [121]	Encoder–Decoder [121]	990 MB [123] and 5.1 GB [122]	https://drive.google.com/file/d/1tZitaNRDaIDK1MPOaQJJn5CivnEIKMnB/view (accessed on 1 January 2023)
2021	KoDF [124]	62,166 (V) [124]	175,776 (V) [124]	237,942 (V) [124]	Face reenactment, Face swap [124]	High [124]	MP4 [125]	1920 × 1080 (initial), 512 × 512 (reduced) [124]	Video: FaceSwap, DeepFaceLab, FSGAN, FOMM; Audio: ATFHP, Wav2Lip [124]	2.6 TB [125]	https://deepbrainai-research.github.io/kodf/
2021	Open Forensics [126]	45,473 (I) [126]	70,325 (I) [126]	115,325 (I) [126]	Face Swap [126]	Low and High (C)	JPG [126]	256 × 256 (low), 512 × 512 (high) [126]	GAN, Poisson blending [126]	56.4 GB [127]	https://zenodo.org/records/5528418 (accessed on 1 January 2024)

Table 12. Comparison and key characteristics of existing face image and video deepfake datasets. I = Images; V = Videos; C = Value or Visual Quality estimated by authors.

Year	Dataset	No. of Real Samples	No. of Fake Samples	Total No. of Samples	Types of Deepfakes	Visual Quality	Format	Resolution	Methods Used to Generate Deepfake	Size of Dataset	Link to Download the Dataset
2021	Perception Synthetic Faces [128]	150 (I) [128]	150 (I) [128]	300 (I) [128]	Entire face synthesis [128]	High (C)	JPG [129]	512 × 512 [129]	PGGAN, StyleGAN, StyleGAN2 [128]	24.5 MB [129]	https://drive.google.com/drive/folders/1d7JhLnXu7r5fm2uZs4EyjGLrkwSgFdYB (accessed on 1 January 2024)
2021	SR-DF [23]	1000 (V) [23]	4000 (V) [23]	5000 (V) (C)	Face swapping, Facial reenactment [23]	High [23]	-	320 × 320 [23]	Face swapping: FS-GAN, FaceShifter; Facial reenactment: First-order-motion, IcFace [23]	-	https://github.com/wangjk666/M2TR-Multi-modal-Multi-scale-Transformers-for-Deepfake-Detection (accessed on 22 April 2024)
2021	Video-Forensics-HQ [130]	-	1737 (V) [130]	1737 (V) [130]	Entire Facial Synthesis, Manipulation [130]	High [130]	MP4 [131]	1280 × 720 [130]	Deep Video Portraits (DVP) [130]	13.5 GB [131]	https://nextcloud.mpi-klsb.mpg.de/index.php/s/EW9bCwCPisfFpww (accessed on 1 January 2024)

Table 13. Comparison and key characteristics of existing face image and video deepfake datasets. I = Images; V = Videos; C = Value or Visual Quality estimated by authors.

Year	Dataset	No. of Real Samples	No. of Fake Samples	Total No. of Samples	Types of Deepfakes	Visual Quality	Format	Resolution	Methods Used to Generate Deepfake	Size of Dataset	Link to Download the Dataset
2021	VideoSham [132]	413 (V) [132]	413 (V) [132]	826 (V) [132]	Face swapping, Face reenactment [132]	High (C)	MP4 [133]	1920 × 1080 [133]	Attack 1(Adding an entity/subject): Adobe Photoshop; Attack 2 (Removing an entity subject): AfterEffects; Attack 3 (Background/Color Change); Attack 4 (Text Replaced/Added); Attack 5 (Frames Duplication/Removal/Dropping); Attack 6 (Audio Replaced) [132]	Trimmed and manipulated: 5.2 GB [134]	https://github.com/adobe-research/VideoSham-dataset (accessed on 1 January 2024)
2021	WPDD [135]	946 (V) [135]	320,499(I) [135]	320,499 (I) + 946 (V) [135]	Face swap, seven face manipulations [135]	Medium and High (C)	MP4	720 × 1280 (Frame), 224 × 224 (Face) [135]	iface and Faceapp	315 GB	-

Table 14. Comparison and key characteristics of existing face image and video deepfake datasets. I = Images; V = Videos; C = Value or Visual Quality estimated by authors.

Year	Dataset	No. of Real Samples	No. of Fake Samples	Total No. of Samples	Types of Deepfakes	Visual Quality	Format	Resolution	Methods Used to Generate Deepfake	Size of Dataset	Link to Download the Dataset
2022	CDDB [136]	-	842,140 (I) [136]	842,140 (I) [136]	Face swap, Face reenactment, Face Synthesis [136]	Medium (C)	PNG [137]	256 × 256 [137]	GAN models: ProGAN, StyleGAN, BigGAN, CycleGAN, GauGAN, StarGAN; Non-GAN models: Glow, CRN, IMLE, SAN, Deepfake, Face2Face, Faceswap, Neural Texture; Unknown model: WhichFaceReal, WildDeepFake [136]	9.6 GB [137]	https://drive.google.com/file/d/1NgB8ytBMFBFwyXJQvdVT_yek1EaaEHrg/view (accessed on 1 January 2024)
2022	CelebV-HQ [138]	-	35,666 (V) [138]	35,666 (V) [138]	Attribute editing [138]	High [138]	MP4 [139]	512 × 512 [138]	VideoGPT, MoCoGAN-HD, DIGAN, StyleGAN-V [138]	38.5 GB [139]	https://pan.baidu.com/s/1TGzOwUcXsRw72l4gaWre_w?pwd=pg71#list/path=%2F (accessed on 1 January 2024)

Table 15. Comparison and key characteristics of existing face image and video deepfake datasets. I = Images; V = Videos; C = Value or Visual Quality estimated by authors.

Year	Dataset	No. of Real Samples	No. of Fake Samples	Total No. of Samples	Types of Deepfakes	Visual Quality	Format	Resolution	Methods Used to Generate Deepfake	Size of Dataset	Link to Download the Dataset
2022	DeePhy [140]	100 (V) [140]	5040 (V) [140]	5140 (V) (C)	Face reenactment, Face swap [140]	High (C)	MPEG4.0 [140]	720 × 720 [140]	FS-GAN, FaceShifter, FaceSwap [140]	26 GB [141]	https://drive.google.com/file/d/1xbFOITBiYe74Oo5_5jysYpprcc3iLFDW/view (accessed on 31 December 2023)
2022	DFDM [142]	-	6450 (V) [142]	6450 (V) [142]	Face swap [142]	High [142]	MPEG4.0 [142]	Average: ∼ 690 × 458 (C)	Faceswap, Lightweight IAE, Dfaker, DFL-H128 [142]	159.59 GB [143]	https://drive.google.com/drive/folders/1aXxeMdA2qwjDytyIgr4CBMVy4pAWizdX (accessed on 1 January 2024)
2022	FakeDance [144]	99 (V) [144]	99 (V) [144]	198 (V) (C)	Whole Body reenactment [144]	Low and High [144]	MP4 [145]	1024 × 512 × 3 [144]	Everybody Dance Now (Video synthesis algorithm), GAN, pix2pixHD [144]	32.2 GB [145]	https://drive.google.com/drive/folders/1IoMA0kEx1IJJVEK0XRR4uPoZtACm6FwC (accessed on 1 January 2024)

Table 16. Comparison and key characteristics of existing face image and video deepfake datasets. I = Images; V = Videos; C = Value or Visual Quality estimated by authors.

Year	Dataset	No. of Real Samples	No. of Fake Samples	Total No. of Samples	Types of Deepfakes	Visual Quality	Format	Resolution	Methods Used to Generate Deepfake	Size of Dataset	Link to Download the Dataset
2022	FMFCC-V [146]	Short version: 44,290 (V), long version: 83 (V) [146]	Short version: 38,102 (V), long version: 192 (V) [146]	Short version: 82,392 (V) [146], long version: 275 (V) (C)	Face swap [146]	High (C)	MP4 [147]	480, 720, 1080 [146]	Faceswap, faceswap-GAN, DeepFaceLab, Recycle-GAN [146]	Long video version: 80.5 GB, Short video version: 411 GB [147]	https://github.com/iiecasligen/FMFCC-V (accessed on 1 January 2024)
2022	GBDF [148]	2500 (V) (C)	10,000 (V) [148]	12,500 (V) (C)	Identity Swap, Expression swap [148]	High (C)	MP4 [149]	256 × 256 [148]	Identity Swapping: FaceSwap, FaceSwap-Kowalski, FaceShifter, Encoder–Decoder; Expression swapping: Face2Face and Neural-Textures [148]	1 TB	https://github.com/aakash4305/~GBDF/releases/tag/v1.0 (accessed on 1 January 2024)
2022	LAV-DF [150]	36,431 (V) [150]	99,873 (V) [150]	136,304 (V) [150]	Face reenactment, Voice Reenactment, Transcript manipulation [150]	High (C)	MP4 [151]	224 × 224 [150]	SV2TTS, Wav2Lip [150]	23.8 GB [151]	https://drive.google.com/file/d/1-OQ-NDtdEyqHNLaZU1Lt9Upk5wVqfYJw/view (accessed on 31 December 2023)

Table 17. Comparison and key characteristics of existing face image and video deepfake datasets. I = Images; V = Videos; C = Value or Visual Quality estimated by authors.

Year	Dataset	No. of Real Samples	No. of Fake Samples	Total No. of Samples	Types of Deepfakes	Visual Quality	Format	Resolution	Methods Used to Generate Deepfake	Size of Dataset	Link to Download the Dataset
2022	SFHQ [152]	-	425,258 (I) [152]	425,258 (I) [152]	Entire face synthesis [152]	High (C)	JPG, PNG [152]	1024 × 1024 [152]	StyleGAN2 [152]	Part 1: 15 GB, Part 2: 15 GB, Part 3: 23 GB, Part 4: 24 GB [152]	Part 1: https://www.kaggle.com/datasets/selfishgene/synthetic-faces-high-quality-sfhq-part-1; Part 2: https://www.kaggle.com/datasets/selfishgene/synthetic-faces-high-quality-sfhq-part-2; Part 3: https://www.kaggle.com/datasets/selfishgene/synthetic-faces-high-quality-sfhq-part-3; Part 4: https://www.kaggle.com/datasets/selfishgene/synthetic-faces-high-quality-sfhq-part-4 (accessed on 1 January 2024)
2022	TrueFace [153]	70,000 (Pre social); 30,000 (Post social) (I) [153]	80,000 (Pre social); 30,000 (Post social) (I) [153]	150,000 (Pre social), 60,000 (Post social) (I) [153]	Face synthesis [153]	High (Quality factor = 87) [153]	JPG [153]	1024 × 1024 (initial), 720 × 720 (resized) [153]	StyleGAN, StyleGAN2 [153]	212 GB [154]	https://drive.google.com/file/d/1WgBrmuKUaLM3YT_5bSgyYUgIUYI_ghOo/view (accessed on 1 January 2024)
2022	ZoomDF [155]	400 (I) [155]	400(V) [155]	400 (I), 400 (V) (C)	Motion manipuation [155]	High (C)	MP4 (C)	-	Avatarify based on First-Order Motion Model (FOMM) [155]	-	-

Table 18. Comparison and key characteristics of existing face image and video deepfake datasets. I = Images; V = Videos; C = Value or Visual Quality estimated by authors.

Year	Dataset	No. of Real Samples	No. of Fake Samples	Total No. of Samples	Types of Deepfakes	Visual Quality	Format	Resolution	Methods Used to Generate Deepfake	Size of Dataset	Link to Download the Dataset
2023	AV-Deepfake 1M [65]	286,721 (V) [65]	860,039 (V) [65]	1,146,760 (V) [65]	Face reenactment, Voice cloning, Transcript manipulation [65]	High [65]	MP4 [156]	224 × 224	TalkLip, VITS, YourTTS, ChatGPT [65]	∼400 GB	https://monashuni-my.sharepoint.com/:f:/g/personal/zhixi_cai_monash_edu/EgeT8-G5RPdLnHqVw33ePRUBeqfxt6Ighe3CmIWKLLQWLQ?e=Sqf7n (accessed on 25 April 2024)
2023	DETER [157]	38,996 (I) [157]	300,000 (I) [157]	338,996 (I) (C)	Face swapping, inpainting, attribute editing [157]	High [157]	JPG [158]	0 × 300; 300 × 512; 512 × 1024; 1024 × 1536; 1536 × 2048; > 2048 [157]	GAN-based models: E4S, MAT; Diffusion Models: DiffSwap, DiffIR [157]	-	https://deter2024.github.io/deter/ (accessed on 1 January 2024)
2023	DFFMD [159]	1000 (V) [159]	1000 (V) [159]	2000 (V) [159]	Facial reenactment [159]	Medium [159]	MP4 [160]	256 × 256 [160]	First-Order Motion model (FOMM) [159]	10 GB [160]	https://www.kaggle.com/datasets/hhalalwi/deepfake-face-mask-dataset-dffmd (accessed on 1 January 2024)
2023	DF-Platter [161]	764 (V) [161]	132,496 (V) [161]	133,260 (V) [161]	Face reenactment, Face swap [161]	High/Low [66], Average brisque score: 43.25 [161]	MPEG4.0 [161]	720 (High resolution), 360 (Low resolution) [161]	FSGAN, FaceSwap, FaceShifter [161]	417 GB [161]	https://drive.google.com/drive/folders/1GeR-a2LfcMkcY6Qzpv2TP8utLtYFBmTs (accessed on 31 December 2023)
2023	eKYC-DF [162]	760 (V) [162]	228,000(V) [162]	228,760 (V) [162]	Face swap [162]	High [162]	-	$512 × 512 and 224 × 224 [162]	SimSwap, FaceDancer, SberSwap [162]	eKYC-DF: >1.5 TB; eKYC-6K: >750 GB	https://github.com/hichemfelouat/eKYC-DF (accessed on 1 January 2024)
2023	IDForge [163]	79,827+, 214,438 (reference dataset) (V) [163]	169,311(V) [163]	463,576 (V) [163]	Face swapping, transcript manipulation, audio cloning/ manipulation [163]	High [163]	-	1280 × 720 [163]	Video: Insight-Face, SimSwap, InfoSwap, Wav2Lip; Audio: TorToiSe, RVC, audio shuffling ; Text: GPT-3.5, text shuffling [163]	∼600 GB	-

Table 19. Comparison and key characteristics of existing face image and video deepfake datasets. I = Images; V = Videos; C = Value or Visual Quality estimated by authors.

Year	Dataset	No. of Real Samples	No. of Fake Samples	Total No. of Samples	Types of Deepfakes	Visual Quality	Format	Resolution	Methods Used to Generate Deepfake	Size of Dataset	Link to Download the Dataset
2023	PolyGlotFake [164]	766 (V) [164]	14,472 (V) [164]	15,238 (V) [164]	Attribute manipulation (lip-sync), text-to-speech, voice cloning [164]	High [164]	-	1280 × 720 [164]	Audio manipulation: Bark+FreeVC, Micro-TTS+FreeVC, XTTS, Tacotron+F-reeVC, Vall-E-X; Video manipulation: VideoRetalking, Wav2Lip [164]	-	https://github.com/tobuta/PolyGlotFake (accessed on 14 June 2024)
2023	Retouching FFHQ [165]	58,158 (I) [165]	652,568 (I) [165]	710,726 (I) (C)	Face retouching [165]	High [165]	JPG, PNG [165]	1024 × 1024 (initial), 512 × 512 (final) [165]	Using APIs: Megvii, Alibaba, Tencent [165]	363.96 GB [166]	https://drive.google.com/drive/folders/194Viqm8Xh8qleYf66kdSIcGVRupUOYvN (accessed on 18 April 2024)
2023	RWDF-23 [167]	-	2000 (V) [167]	2000 (V) [167]	Face swapping [167]	High [167]	-	> 1000 × 1000 [167]	DeepFaceLab, DeepFaceLive, FOMM, SimSwap, FacePlay, Reface, Deepfake Studio, FaceApp, Revive, LicoLico, Fakeit; Deep-Faker, Deep-FakesWeb, Deepcake.io, DeepFaker Bot, Revel.ai [167]	-	-

Table 20. Comparison and key characteristics of existing face image and video deepfake datasets. I = Images; V = Videos; C = Value or Visual Quality estimated by authors.

Year	Dataset	No. of Real Samples	No. of Fake Samples	Total No. of Samples	Types of Deepfakes	Visual Quality	Format	Resolution	Methods Used to Generate Deepfake	Size of Dataset	Link to Download the Dataset
2023	SPRITZ-PS [168]	20,000 (I) [168]	20,000 (I) [168]	40,000 (I) [168]	Entire face synthesis, iris reconstruction [168]	Low to Medium (C)	JPG Face: $> 2000\times > 2000$; Iris: 256×256 [169]	StyleGAN2, ProgressiveGAn, StarGAN [168]	17.84 GB [169]	https://ieee-dataport.org/documents/spritz-ps-validation-synthetic-face-images-using-large-dataset-printed-documents (accessed on 1 January 2024)	
2024	DeepFaceGen [170]	463,583 (I), 313,407 (V) [170]	350,264 (I), 423,548 (V) [170]	350,264 (I), 423,548 (V) [170]	Entire Face Synthesis, Face Swapping, Face Reenactment, Attribute Manipulation [170]	High [170]	JPG, PNG, MP4 [171]	-	FaceShifter, FSGAN, DeepFakes, BlendFace, DSS, SBS, MMReplacement, SimSwap, Talking Head Video, ATVG-Net, Motion-cos, FOMM, StyleGAN2, MaskGAN, StarGAN2, SC-FEGAN, DiscoFaceGAN, OJ, SD1, SD2, SDXL, Wenxin, Midjourney, DF-GAN, DALL·E, DALL·E 3, AnimateDiff, AnimateLCM, Hotshot, Zeroscope, MagicTime, Pix2Pix, SDXLR, VD [170]	491.1 GB [171]	https://github.com/HengruiLou/DeepFaceGen (accessed on 20 June 2024)

Table 21. Comparison and key characteristics of existing face image and video deepfake datasets. I = Images; V = Videos; C = Value or Visual Quality estimated by authors.

Year	Dataset	No. of Real Samples	No. of Fake Samples	Total No. of Samples	Types of Deepfakes	Visual Quality	Format	Resolution	Methods Used to Generate Deepfake	Size of Dataset	Link to Download the Dataset
2024	DF40 [172]	52,590 (V) [172]	0.1 M+ (I), 1 M+ (V) [172]	0.1 M+ (I), 1 M+ (V) [172]	Face Swapping, Face Reenactment, Entire Face Synthesis, Attribute Manipulation [172]	High [172]	-	-	FSGAN, FaceSwap, SimSwap, InSwapper, BlendFace, UniFace, MobileSwap, e4s, FaceDancer, DeepFaceLab, FOMM, FS_vid2vid, Wav2Lip, MRAA, OneShot, PIRender, TPSMM, LIA, DaGAN, SadTalker, MCNet, HyperReenact, HeyGen, VQGAN, StyleGAN2, StyleGAN3, StyleGAN-XL, SD-2.1, DDPM, RDDM, PixArt-α, DiT-XL/2, SiT-XL/2, MidJounery6, WhichisReal, CollabDiff, e4e, StarGAN, StarGANv2, StyleCLIP [172]	-	-

2.5. EBV (Eye Blinking Video) Dataset [55]

Within this dataset, there are 50 authentic videos as well as 49 counterfeit videos, focusing on deepfakes of the face-swapping type. It employs the DeepFake algorithm [55] for generation. The videos have a resolution of 224 × 224 pixels, characterized by low visual quality. The dataset is 1.01 GB in size.

2.6. Faceforensics [57]

This database was introduced in 2018. It comprises 1004 real and 1004 deepfake videos. The source original videos were acquired from YouTube. The videos tagged with 'face', 'newscaster' or 'newsprogram' on YouTube and the YouTube-8M dataset [173] were chosen for this dataset. The dataset focuses on two types of deepfakes (i.e., Self-Reenactment and Source-to-target transfer) using the Face2Face technique. In terms of visual quality, it is high quality and of MP4 format with a resolution of 128 × 128 pixels. The dataset has been segmented into 704, 150, and 150 videos for training, validation, and testing, respectively. The dataset consists of 130 GB of lossless compressed videos and 3.5 TB of raw videos.

2.7. FFW (Fake Faces in the Wild) [61]

This database, established in 2018, comprises a total of 150 fake videos. The dataset focuses primarily on face replacement and image tampering. The videos are formatted in MP4 and AVI, and exhibit a resolution of 480 pixels or higher. The manipulation techniques used are CGI (Computer-Generated Imagery), GANs (generative adversarial networks), and FakeApp mobile application.

2.8. HOHA (Hollywood Human Actions)-Based [63]

The HOHA-based dataset was introduced in 2018 and comprises 300 real and 300 forged videos. This dataset primarily focuses on face swapping and used the encoder–decoder approach for deepfake generation. The videos are provided in MP4 format with a resolution of 299 × 299 pixels, characterized by low-to-medium visual quality.

2.9. UADFV [64]

UADFV is a deepfake dataset created in 2018 with 98 samples, evenly distributed between 49 deepfake and 49 real videos. The dataset's primary focus is on face swap deepfakes. It has low visual quality with a resolution of 294 × 500 pixels. The FakeApp mobile application was used to generate the deepfakes. This database has a size of 146 MB.

2.10. Celeb-DF [68]

This is another deepfake dataset featuring 6229 samples, including 590 real and 5639 fake samples. The primary focus of the dataset is on face reconstruction, accomplished using an improved deepfake synthesis algorithm. The visual quality of the deepfakes is high. The videos in the dataset are in MPEG4.0 format and have a resolution of 256 × 256 pixels. The dataset was generated using the original deepfake synthesis algorithm. The dataset has a size of 9.3 GB. The real videos acting as sources were derived from publicly accessible YouTube clips, showcasing 59 celebrities with diverse gender identities, age groups, and ethnic backgrounds. In the genuine videos, 56.8% of the subjects are male while 43.2% are female. The distribution of subjects is as follows: 6.4% are under the age of 30, 28.0% are in their 30s, 26.6% are in their 40s, 30.5% are in the 50–60 age range, and 8.5% are aged 60 and above. In terms of ethnicity, 5.1% are Asians, 88.1% are Caucasians, and 6.8% are African Americans.

2.11. DeepFakeDetection (Google and Jigsaw Dataset) [70]

This dataset was published in 2019 by Google and Jigsaw. It includes 3068 fake videos and 363 real videos. The deepfake type used in the dataset is face swap. The dataset includes videos manipulated with DeepFakes (DF) [174], Face2Face (F2F) [59], FaceSwap (FS) [175], and NeuralTextures (NT) [176], and is part of the FaceForensics benchmark.

The source videos have the following sizes based on their compression rate factors: Raw/0: 200 GB, C23: 3 GB, and C40: 400 MB. The manipulated videos have the following sizes: Raw/0: 1.6 TB, C23: 22 GB, and C40: 3 GB.

2.12. DEFACTO [72]

The DEFACTO deepfake dataset has a total of 299,000 fake images. For faces, the dataset focuses on face swapping and face morphing. It used techniques like copy–move, splicing, object-removal, and morphing. The copy–move technique was used for duplication of an element within the image. For splicing, one portion of an image is copied and pasted onto another image. In object-removal, an object is removed from the image using inpainting algorithms. Finally, morphing consists of warping and blending two images together. For each forgery, postprocessing may be applied (rotation, scaling, contrast, etc.). The visual quality of the deepfakes in DEFACTO is low. The images are encoded in the TIF (Tagged Image File Format) and JPG formats and have resolutions ranging from 240 × 320 pixels to 640 × 640 pixels. It uses the FaceSwap technique to generate deepfakes. The dataset has a size of 121 GB.

2.13. DFFD (Diverse Fake Face Dataset) [74]

This dataset was made available in 2019 and is made up of 1000 real videos and 58,703 genuine images, alongside 3000 deepfake videos and 240,336 deepfake images. The deepfake variants encompass identity swap, expression swap, attribute manipulation, and entire face synthesis using frameworks such as FaceSwap [174], Deepfake [174], Deep Face Lab [177], Face2Face [59], FaceAPP [10], StarGAN [178], PGGAN (Progressive Growing of GANs) [50], and StyleGAN [78]. The visual quality is low and high. The size of the dataset is 18.73 GB.

2.14. FaceForensics++ [76]

FaceForensics++ stands out as a prominent dataset for deepfake detection. It is an extension of the Faceforensics [57] database. It is made up of 1000 real and 4000 manipulated videos. Multiple tools and techniques were employed to generated four subsets of deepfakes, i.e., DeepFakes (DF) [174], Face2Face (F2F) [59], FaceSwap (FS) [175], and NeuralTextures (NT) [176]. The visual qualities are mainly low and high. The dataset utilizes uncompressed and H264 compressed format, i.e., H.264 encoding with CRF values of 0, 23, and 40 for the MP4 format videos. The deepfake detector's performance can be assessed on both compressed and uncompressed videos. The dataset offers a range of resolutions: 854 × 480 for VGA, 280 × 720 for HD, and 920 × 1080 for Full HD. Lack of lip-sync deepfakes and color inconsistencies can be easily observed in the dataset. This dataset contains 60% male and 40% female samples. The original videos have a total size of 38.5 GB. The h264 compressed videos have the following sizes based on their compression rate factors: Raw/0: 500 GB, 23: 10 GB, and 40: 2 GB. Moreover, all the raw extracted images in PNG format have a total size of 2 TB.

2.15. FFHQ (FlickrFaces-High Quality) [78]

The FFHQ (Flickr-Faces-HQ) dataset was created in 2019 with 70,000 real and 140,000 fake images, totaling 210,000 images. The dataset focuses on high-resolution synthetic and attribute-manipulated images stored in PNG format, each with a resolution of 1024 × 1024 pixels. The images were generated using StyleGAN technology. It has an extensive size of 2.56 TB.

2.16. WhichFaceReal [80]

This dataset was established in 2019 and contains of 70,000 deepfake images, focusing primarily on face synthesis deepfakes. Stored in JPG, the samples maintain a resolution of 1024 × 1024 pixels. The dataset used the StyleGAN [11] technique to generate manipulated images.

2.17. DeeperForensics-1.0 [81,82]

DeeperForensics-1.0 is a deepfake dataset created in 2020 containing 59,475 videos, including 48,475 real and 11,000 fake videos. The primary focus of this dataset is on face swapping. The visual quality of the deepfakes in DeeperForensics-1.0 is high and the videos are encoded in MP4 format. The resolution of the content is 1920 × 1080 pixels. The generation process involves the use of DF-VAE (DeepFake Variational Auto Encoder). The dataset has a size of 281.35 GB. It includes 45 females and 55 males, spanning across 26 countries. The age spectrum of the subjects varies between 20 and 45 years, mirroring the prevalent age range in real-world videos. The videos contain eight expressions: neutral, happy, surprise, sad, angry, disgust, contempt, and fear.

2.18. DFDC (Deepfake Detection Challenge) [85]

DFDC is a public dataset created in 2020, comprising 128,154 videos. This dataset includes 104,500 fake and 23,654 real videos. It contains face reenactment and face swap types of deepfakes. The deepfakes exhibit a high level of visual quality, and videos are in the MP4 format. The dataset offers two resolutions, with images sized at 128 × 128 pixels (DF-128) and 256 × 256 pixels (DF-256). Multiple tools and techniques were employed in the generation process, including DFAE (Deepfake Autoencoder), MM/NN (morphable-mask/nearest-neighbors model), NTH (Neural Talking Heads) [179], FSGAN (Face Swapping GAN) [180], StyleGAN [11], Refinement, and Audio Swaps [181]. This dataset has a size of 471.84 GB.

2.19. DMD (Deepfake MUCT Dataset) [87]

This dataset was created in 2020. It contains 751 original images and 8857 manipulated images, forming a dataset of 9608 images in total. The dataset focuses on deepfake types such as face swapping, gender conversion, face morphing, and smile manipulation. All images have a resolution of 640 × 480 pixels. The FakeApp mobile application was used to generate the deepfakes. It has a total size of 1.42 GB.

2.20. FaceShifter [88]

This dataset is composed of 1000 real videos and 10,000 fake videos. It is primarily centered around the face swapping deepfake type. It contains deepfake images generated using the FaceShifter technique with a resolution of 256 × 256 pixels. The videos have a high visual quality.

2.21. FakeET [89]

Introduced in 2020, this dataset includes 331 authentic videos and 480 manipulated (fake) videos. This dataset focuses on face swap. The videos are presented in high visual quality at 1920 × 1080 pixels and are in MP4 format. This dataset was generated using the Google/Jigsaw dataset. This dataset has a size of 25.9 GB.

2.22. FFIW 10K (Face Forensics in the Wild) [91]

FFIW was introduced in 2020, featuring 10 K real and 10 K fake videos, totaling 20,000 videos. The deepfake type used in this dataset is face swapping. The videos in FFIW have a resolution of 480 pixels and above. The dataset uses face manipulation technologies like DeepFaceLab [182], FS-GAN [180], and FaceSwap [183]. The dataset is split into a training set of 17 gigabytes and a testing set with a size of 4.1 GB. The videos are in MP4 format with high visual quality.

2.23. iFakeFaceDB [94]

This dataset was also introduced in 2020, focusing on entire face synthesis. It includes a total of 87,000 fake images. The images within iFakeFaceDB are stored in jpg format and have a resolution of 224 × 224 pixels. The high-resolution and visual quality of the

images ensures detailed and visually rich synthetic faces. The dataset is generated using StyleGAN [11] and GANPrintR approaches. This dataset has a size of 1.4 gigabytes.

2.24. RFFD (Real and Fake Face Detection) [97]

The dataset is composed of 1081 genuine images and 960 manipulated images in training. The dataset was focused on generating attribute-manipulated deepfakes using Photoshop. The videos are standardized at a resolution of 600 × 600 pixels. The dataset size is 431 megabytes. The images are in the format of JPG with a medium-to-high visual quality.

2.25. UIBVFED [98]

UIBVFED (User-Independent Blendshape-based Video Facial Expression Dataset) contains 640 fake images. This dataset used the blendshapes technique and autodesk character generator tool to generate attribute manipulation and entire face synthesis-based deepfakes. It has a medium-to-high visual quality. The dataset contains the images in the PNG format. The resolution for each image in UIBVFED is set at 750 × 133 pixels. The dataset has a size of 652 megabytes.

2.26. WildDeepfake [100]

WildDeepfake is a unique deepfake dataset created in 2020 with a total of 7314 videos, comprising 3805 real and 3509 fake videos. This dataset was collected from various sources on Internet. The visual quality of the deepfakes is high. It is acknowledged as a challenging dataset for deepfake detection. The images are in the PNG format with a resolution of 224 × 224 pixels. The dataset has a size of 67.8 gigabytes.

2.27. YouTube-DF [102]

The dataset includes 98 authentic videos that were utilized to generate 79 deepfake videos and 98 faceswaps. It mainly focuses on face swapping and has low and high visual qualities. The videos are formatted in MP4 at 256 × 256 pixels. DeepFaceLab was the technology employed for face swapping in this dataset.

2.28. DeepFake MNIST+ [103]

Crafted in 2021, this dataset comprises 20,000 videos, meticulously balanced with 10,000 authentic videos and 10,000 manipulated (fake) videos. This dataset primarily focuses on image animation. The visual quality of the deepfakes in DeepFake MNIST+ is high. The videos are in MP4 format and have a resolution of 256 × 256 pixels. The generation process involves the use of the FOMM (First-Order Motion Model) technique. It has a size of 2.21 GB. The videos contain ten actions: open mouth, blink, yawn, left slope head, right slope head, nod, surprise, embarrassment, and look up and smile.

2.29. DeepStreets [106]

This dataset includes a total of 1200 video samples, evenly distributed with 600 authentic videos and 600 manipulated (fake) videos. The type of deepfake used in dataset is video-to-video synthesis. It has low and high visual qualities, and utilizes Vid2vid and Wcvid2vid technologies [184,185] to generate deepfakes. The videos are in MP4 format with a resolution of 512 × 1024 pixels. The total size of the dataset is 24.1 GB.

2.30. DFGC-21 (DeepFake Game Competition) [108]

This dataset was introduced in 2021. It focuses on face swapping deepfake types, encompassing 1000 real images and $N \times 1000$ fake images. The images within DFGC-21 are in PNG format and feature a resolution of 300 × 300 pixels. It uses various face manipulation techniques such as FaceShifter [88], FaceController [186], FaceSwap [183], FOMM [105], and FSGAN [180]. The samples' visual quality is high and the total dataset size is 5.18 GB.

2.31. DF-Mobio [110]

This is a gender-diverse dataset released in 2021. It contains a total of 46,496 deepfake videos, including 31,950 real and 14,546 fake video samples focusing on identity swap. The videos are characterized by a pixel resolution of 256 × 256 and were generated using GAN technology. The dataset consists of bimodal (audio and video) data taken from 150 people. The dataset has a female–male ratio of nearly 1:2 (99 males and 51 females) and was collected from August 2008 until July 2010 at six different sites from five different countries. The videos are in MP4 format with high visual quality, and the dataset has a size of 170.1 GB.

2.32. DF-W (DeepFake Videos in the Wild) [112]

This dataset consists of 1869 manipulated (fake) videos. The dataset focuses on face swapping. The videos are collected from YouTube, Reddit, and Bilibili by searching for videos with keywords 'face swap'. The videos in DF-W are formatted in MP4 and are available in two resolutions: a lower resolution of 360 × 360 pixels and a higher resolution of 2560 × 1080 pixels. This dataset has a compact size of 31.55 gigabytes.

2.33. FaceSynthesis [114]

This dataset has a collection of 100,000 fake images. The images have a resolution of 512 × 512 pixels. The synthesis techniques involved are associated with VFX (Visual Effects). The images are in PNG format with high visual quality. The total dataset size is 31.8 gigabytes.

2.34. FakeAVCeleb [116]

This dataset comprises 20,000 videos, with 500 authentic videos and 19,500 manipulated (fake) videos. The generated deepfakes encompass face-swapping, facial reenactment, and voice cloning. The videos in FakeAVCeleb are presented in MP4 format with high visual quality and a resolution of 224 × 224 pixels. For video manipulations, technologies such as FSGAN [180], Wav2Lip [187], and FaceSwap [188] were utilized, while SV2TTS (Speaker Verification to Text-to-Speech) [20] was used for audio manipulations. This dataset has a size of 6 gigabytes.

2.35. ForgeryNet [119]

Introduced in 2021, this extensive deepfake dataset comprises a total of 2,896,062 images and 221,247 videos. It is characterized by a balance of 1,438,201 real images and 1,457,861 manipulated (fake) images, as well as 121,617 fake videos and 99,630 real videos. The dataset encompasses various face manipulation techniques like face editing, face reenactment, face transfer, face swap, and face stacked manipulation. The visual quality of the deepfakes in ForgeryNet varies for both images and videos (high and low). The dataset's images and videos are encoded in the JPG and MP4 formats, exhibiting a diverse range of resolutions spanning from 240 pixels to 1080 pixels. The generation process involves the use of multiple techniques and tools, such as FSGAN [180], FaceShifter [88], BlendFace, FOMM Replacement [105], MASKGAN [189], StarGAN [190], StyleGAN [11], ATVG-Net (audio transformation and visual generation network) [191], and SC-FEGAN [192]. The total size of the dataset is 496.23 GB.

2.36. HiFiFace (High Fidelity Face Swapping) [121]

The HiFiFace database encompasses 1000 synthetic videos derived from FaceForensics++, aligning precisely with the target and source pair configurations in FaceForensics++. In addition, 10,000 frames sourced from FaceForensics++ videos are included in the database, providing a robust foundation for quantitative assessment. The face swapping was performed using an architecture composed of three components, i.e., a 3D shape-aware identity extractor, a semantic facial fusion module, and an encoder–decoder structure.

The samples have two resolutions, i.e., 256 × 256 and 512 × 512. The database is 900 MB in size.

2.37. KoDF (Korean DeepFake Detection Dataset) [124]

Established in 2021, this deepfake dataset encompasses a total of 237,942 videos, consisting of 62,166 authentic videos and 175,776 manipulated (fake) videos. The dataset focuses on both face reenactment and face swap deepfakes. The visual quality of the deepfakes in KoDF is high. The videos have an initial resolution of 1920 × 1080 pixels, which is later reduced to 512 × 512 pixels. Various tools and techniques, including FaceSwap [174], DeepFaceLab [182], FSGAN [180], FOMM (First-Order Motion Model) [105], ATFHP (Audio-driven Talking Face Head Pose) [193], and Wav2Lip [187], were used in the deepfake generation process. The surveyed population was predominantly aged between 20 and 39, making up 77.21% of the respondents. Gender distribution is nearly equal, with females at 50.87% and males at 49.13%, reflecting a balanced demographic profile. The dataset has a size of 2.6 TB.

2.38. OpenForensics [126]

This dataset consists of 45,473 authentic images and 70,325 deepfake images. It specifically concentrates on generating face swap deepfake images. The images in OpenForensics are presented in JPG format and have resolutions of 256 × 256 (low) and 512 × 512 (high) pixels with low and high visual qualities. The dataset synthesizes images using advanced generative techniques, such as GAN (Generative Adversarial Network) [194,195] and Poisson blending [196]. This dataset has a size of 56.4 gigabytes.

2.39. Perception Synthetic Faces [128]

This dataset was introduced in 2021 and consists of 150 authentic images and 150 manipulated (fake) images. This dataset synthesizes the entire face to generate deepfakes using techniques such as PGGAN [50], StyleGAN [78], and StyleGAN2 [197]. The samples are in jpg format, featuring a resolution of 512 × 512 pixels, and exhibit high visual quality. This dataset has a size of 24.5 megabytes.

2.40. SR-DF (Swapping and Reenactment DeepFake) [23]

Introduced in 2021, this dataset is centered around facial reenactment and face swapping, consisting of 1000 real videos and 4000 manipulated (fake) videos. The videos in SR-DF have a resolution of 320 × 320 pixels. The dataset incorporates a variety of face manipulation techniques, with face swapping employing technologies such as FS-GAN [180] and FaceShifter [88], while facial reenactment involves the First-order motion model and IcFace. The dataset has a high visual quality.

2.41. Video-Forensics-HQ [130]

This dataset was established in 2021 and contains 1737 fake videos. The dataset contains high-quality deepfakes of entire face synthesis and manipulation types. The videos are in MP4 format and have a resolution of 1280 × 720 pixels. The dataset used techniques such as Deep Video Portraits (DVP) [198] to achieve its synthetic images. It has a total size of 13.5 GB.

2.42. VideoSham [132]

This database presents a balanced composition, featuring an equal number of fake and real samples, totaling 826 videos. The deepfake samples are face swap and face reenactment types. This dataset uses spatial and temporal stacks as techniques to generate deepfakes. Spatial attacks (1–4) include adding/removing entities, changing backgrounds, and text manipulation, while temporal attacks (5–6) involve frame and audio modifications. The videos have a resolution of 1920 × 1080 pixels. The videos are in MP4 format with high visual quality. The dataset has a total size of 5.2 GB.

2.43. WPDD (World Politicians Deepfake Dataset) [135]

WPDD was created in 2021 and consists of 946 real videos and 320,499 fake images. The dataset was primarily created using face swap techniques and also incorporates seven distinct face manipulations. The videos within WPDD are in MP4 format and possess resolutions of 224 × 224 pixels and 720 × 1280 pixels. The videos manipulations were created using the iFace and FaceApp applications with medium and high visual qualities. This dataset has a size of 315 gigabytes.

2.44. CDDB (Continual Deepfake Detection Benchmark) [136]

The CDDB dataset was introduced in 2022. It contains 842,140 deepfake images. It mainly focuses on three deepfakes types: face synthesis, face swap, and face reenactment. The dataset is characterized by a variety of GAN models such as ProGAN [50], StyleGAN [78], BigGAN [199], CycleGAN [200], GauGAN (Gaussian GAN) [201], StarGAN [178], non-GAN models such as Glow [202], CRN (Cascaded Refinement Network) [203], IMLE (Implicit Maximum Likelihood Estimation) [204], SAN (Second-order Attention Network) [205], DeepFakes [174], Face2Face [59], FaceSwap [175], and NeuralTextures [176], and unknown models such as WhichFaceReal [80] and WildDeepFake [100]. It has a total size of 9.6 gigabytes. The images, formatted in PNG, have a resolution of 256 × 256 pixels and medium visual quality.

2.45. CelebV-HQ (High-Quality Celebrity Video Dataset) [138]

This database comprises 35,666 fake videos, specifically emphasizing high-quality renditions of celebrity faces. This dataset focuses on attribute manipulation and employs techniques such as VideoGPT [206], MoCoGAN-HD (Motion and Content decomposed GAN for HighDefinition video synthesis) [207], DIGAN (Dynamics-aware Implicit Generative Adversarial Network) [208], and StyleGAN-V [209]. The videos in CelebV-HQ are presented in MP4 format and have a resolution of 512 × 512 pixels. This dataset has a size of 38.5 gigabytes. All clips underwent manual labeling, encompassing 83 facial attributes that span appearance, emotion, and action.

2.46. DeePhy (Deepfake Phylogeny) [140]

DeePhy is a deepfake dataset created in 2022 comprising 5140 videos, with 100 real and 5040 fake videos. The dataset covers multiple deepfakes types, including face reenactment and face swap. The visual quality of the deepfakes in DeePhy is high. The videos are in MPEG4.0 format with a resolution of 720 pixels. The generation process involves the use of the following techniques: FS-GAN [180], FaceShifter [88], and FaceSwap [183]. It has a size of 26 GB. The generated deepfakes are segregated in three categories: (i) utilizing a single technique, (ii) employing two different techniques, and (iii) applying three different techniques.

2.47. DFDM (DeepFakes from Different Models) [142]

The DFDM dataset has a collection of 6450 deepfake videos that focus on face-swapping. The videos in DFDM are of high quality and presented in the MPEG4.0 format, with an average resolution of 690 × 458 pixels. This dataset incorporates various face manipulation techniques like FaceSwap [183], Lightweight [174], IAE, Dfaker [210], and DFL-H128 [182]. It has a size of 159.59 gigabytes.

2.48. FakeDance [144]

This dataset includes 99 real and 99 fabricated videos. It focuses on the whole body reenactment deepfake. It utilizes Everybody Dance Now (a video synthesis algorithm) [211], GAN, and pix2pixHD [212] technologies to generate deepfakes. The videos are in MP4 format with a resolution of 1024 × 512 × 3 pixels and low/high visual quality. The total size of the dataset is 32.2 GB.

2.49. FMFCC-V (Fake Media Forensics Challenge of China Society of Image and Graphics-Video Track) [146]

This dataset was introduced in 2022 and features 42 male and 41 female subjects. The short video version has 44,290 real and 38,102 fake videos, totaling 82,392 videos, whereas the long video version has 83 real videos and 192 fake videos, totaling 275 videos. It uses various face manipulation tools like faceswap [174], faceswapGAN [52], DeepFaceLab [177], Recycle-GAN [213], and Poisson blending. This dataset contains two versions: long and short. The long version is 80.5 GB and the short version is 411 GB in size.

2.50. GBDF (Gender Balanced DeepFake Dataset) [148]

This dataset was introduced in 2022. It is a specialized dataset created to address gender balance in the context of deepfake generation and detection. The dataset comprises 2500 real and 10,000 fake videos. The videos in GBDF are of high quality, presented in MP4 format with a resolution of 256 × 256 pixels. The manipulation techniques used in this dataset are identity swapping (FaceSwap, FaceSwap-Kowalski [183], FaceShifter [88], DeepFakes [174], and Encoder–Decoder) and expression swapping (Face2Face [59] and NeuralTextures) [176]. The dataset is 1 TB in size.

2.51. LAV-DF (Localized Audio Visual DeepFake) [150]

This is a significant deepfake dataset created in 2022 comprising 136,304 videos, with 36,431 real and 99,873 fake videos. The dataset focuses on face reenactment, voice reenactment, and transcript manipulation. The visual fidelity of the deepfakes in LAV-DF is high. The dataset includes both video files (MP4) and associated metadata files (csv). The initial resolution of the images is 224 × 224 pixels. The generation process involves the use of specific tools and techniques, including SV2TTS [20] and Wav2Lip [187]. It has a size of 23.8 GB.

2.52. SFHQ (Synthetic Faces High-Quality) [152]

The SFHQ dataset was created in 2022. It contains 425,258 synthesized images. These images are provided in multiple formats, including jpg and PNG. The resolution for each image is set at 1024 × 1024 pixels. It uses StyleGAN2 [197] to generate deepfakes. The dataset is distributed across four parts, each varying in size: Part 1 (15 GB), Part 2 (15 GB), Part 3 (23 GB), and Part 4 (24 GB).

2.53. TrueFace [153]

This dataset is divided into two parts: pre-social and post-social. The pre-social dataset consists of 70,000 authentic images and 80,000 synthetic images, amounting to a total of 150,000 images. The post-social dataset contains 30,000 real and 30,000 fake images, totaling 60,000 images. It contains images of high visual quality with a quality factor of 87. It uses generation techniques like StyleGAN [78] and StyleGAN2 [197]. The videos are presented at a resolution of 1024 × 1024 pixels initially and 720 × 720 pixels after resizing. The total size of the dataset is 212 GB.

2.54. ZoomDF [155]

This dataset was introduced in 2022 with 400 fake video samples generated from 400 real images. The deepfake videos were generated on the Avatarify framework using the First-Order Motion Model (FOMM) [105] method. The type of deepfake is motion manipulation. The videos are in MP4 format with a high visual quality.

2.55. AV-Deepfake1M [65]

Within this dataset, there are 286,721 legitimate videos alongside 860,039 manipulated videos. It concentrates on the creation of deepfakes through face reenactment, as well as the voice cloning and manipulation of transcripts. It uses TalkLip [214], VITS (Variational Inference with adversarial learning for end-to-end Text-to-Speech) [215], ChatGPT [216] for

transcript manipulation, and YourTTS [217] for manipulation. The samples in this dataset are in MP4 format with a resolution of 224 × 224 pixels and high visual quality. It has a total size of ∼400 GB.

2.56. DETER (DETEcting Edited Image Regions) [157]

DETER is a large-scale, manipulated deepfake database. The dataset comprises 38,996 authentic images and 300,000 forged images. The deepfake images were generated by four state-of-the-art techniques with three distinct editing operations, i.e., face swapping, attribute editing, and inpainting. The face swapping and attribute editing were performed using GANs-based editing for swapping (E4S) [218] and diffusion models (DMs)-based DiffSwap [219]. For inpainting, the manipulation tools employed were the GANs-based Mask-Aware Transformer (MAT) [220] and DMs-based DiffIR [221]. In the process of face swapping and attribute editing, modifications were focused on specific facial regions such as eyes and nose. In contrast, inpainting operations targeted random image regions, effectively eliminating spurious correlations present in prior datasets. Furthermore, meticulous image postprocessing was executed to guarantee that the deepfakes maintain a realistic appearance. The high visual quality of the edited images is evident within the dataset. The image resolutions ranges from $0 - 300 \times 0 - 300$ to 1536×2048 and >2028.

2.57. DFFMD (Deepfake Face Mask Dataset) [159]

This dataset, introduced in 2023, comprises 1000 real and 1000 fabricated videos, resulting in a total of 2000 videos. The dataset focuses on facial reenactment. The technique used in the generation process is FOMM [105]. The videos are in MP4 format with a resolution of 256 × 256 pixels and medium visual quality. The dataset contains videos of 40 subjects—28 male and 12 female. The dataset has a total size of 10 GB.

2.58. DF-Platter [161]

This is a deepfake dataset with a total of 133,260 videos, consisting of 764 real and 132,496 fake videos. The primary focus of this dataset is on face reenactment and face swap. The dataset includes an average BRISQUE score of 43.25, indicating high visual quality. The content is encoded in the MPEG4.0 format and is provided in two resolutions: 720 pixels (high resolution) and 360 pixels (low resolution). The generation process involves the use of various tools, including FSGAN [180], FaceSwap [183], and FaceShifter [88]. It has a size of 417 GB. This database comprises three distinct sets: Set A, Set B, and Set C. Set A encompasses deepfakes featuring a single subject. To generate single-subject deepfakes, a source video and a target video are utilized, each featuring a single subject. Set B encompasses intra-deepfakes, where the faces of one or more subjects within a given video are swapped. In contrast, Set C comprises multi-face deepfakes, wherein the faces in the source videos undergo manipulation to resemble those of celebrities in the target. Set A incorporates deepfakes created using FSGAN and FaceShifter, while Sets B and C include deepfakes generated through all three generation schemes.

2.59. eKYC-DF [162]

This dataset was also introduced in 2023 and contains 228,000 fake video and 760 real video samples. It utilized the SimSwap [222], FaceDancer [223], and SberSwap [224] schemes to yield face-swapped deepfakes. The samples in this dataset are of 512 × 512 and 224 × 224 resolutions. The visual quality of the samples is high. The dataset sizes are estimated to be over 1.7 TB for the eKYC-DF dataset and 750 GB for the eKYC-6K dataset.

2.60. IDForge [163]

This dataset was introduced in 2023. It is composed of 79,827 real samples along with 214,438 samples in the reference dataset as well as 169,311 deepfake samples. It uses various techniques to produce deepfakes like face swapping with lip-syncing (InsightFace [225,226], SimSwap [222], InfoSwap [227], and Wav2Lip [187]), audio cloning/manipu-

lation (TorToiSe [228], RVC [229], and audio shuffling), and transcript manipulation (GPT-3.5 [230] and text shuffling). The samples have a resolution of 1280 × 720 pixels. The visual quality of the samples is high, and the dataset size totals an estimate of ∼600 GB.

2.61. PolyGlotFake [164]

The dataset contains an audiovisual and multilingual deepfake collection, which includes 766 real videos and 14,472 fake videos, amounting to a total of 15,238 videos. The samples are in seven languages, i.e., English, French, Spanish, Russian, Chinese, Arabic, and Japanese. The audio and video manipulations were generated utilizing various text-to-speech, voice cloning, and lip-sync methods. For audio manipulations, Bark+FreeVC [231,232], MicroTTS+FreeVC [233], XTTS [234], Tacotron+FreeVC [235], and Vall-E-X [236] were used. VideoRetalking [237] and Wav2Lip [187] were employed for video manipulations. The quality of the deepfakes is high.

2.62. RetouchingFFHQ (Retouching FlickrFaces-High Quality) [165]

Established in 2023, this dataset consists of 710,726 images, including 58,158 real and 652,568 fabricated images. It primarily concentrates on facial retouching tasks, encompassing skin smoothing, face whitening, face lifting, and eye enlarging. The visual quality of the deepfakes in RetouchingFFHQ is high. The images are initially provided at a resolution of 1024 × 1024 pixels, which is later reduced to 512 × 512 pixels in the final output. The dataset is generated using APIs from major technology companies, including Megvii [238], Alibaba [239], and Tencent [240]. The dataset has a total size of 363.96 GB.

2.63. RWDF-23 (Real-World Deepfake) [167]

The RWDF database was introduced in 2023, encompassing content in English, Korean, Chinese, and Russian languages. The dataset consists of 2000 fake videos sourced from diverse platforms like YouTube, TikTok, Reddit, and Bilibili. These videos are presented with a resolution greater than 1000 × 1000 pixels. The deepfakes were generated using open source frameworks (i.e., DeepFaceLab [177], DeepFaceLive [241,242], FOMM [105], and SimSwap [222]), mobile apps (i.e., FacePlay [243], Reface [244–246], Deepfake Studio [247], FaceApp [10], Revive [248], LicoLico [249], Fakeit [250], and DeepFaker [251]) and commercial products (i.e., DeepFakesWeb [252], Deepcake.io [253,254], DeepFaker Bot [255], and Revel.ai [256]).

2.64. SPRITZ-PS [168]

Within this database, there are a total of 20,000 deepfake images. It focuses on synthetic false images and iris reconstruction using techniques such as StyleGAN2, ProgressiveGAN, and StarGAN. The dataset provides images in jpg format. It has a size of 17.84 gigabytes.

2.65. DeepFaceGen [170]

DeepFaceGen was created in 2024 and contains 463,583 real images, 313,407 real videos, 350,264 forged images, and 423,548 forged videos. In total, it is composed of 813,847 images and 736,955 videos. The facial deepfakes are focused on entire face synthesis, face swapping, face reenactment, and attribute manipulation, which were generated utilizing 34 image or video generation methods. The deepfake generation techniques employed include FaceShifter [88], FSGAN [180], DeepFakes [174], BlendFace [257], MMReplacement [119], DeepFakes-StarGAN-Stack (DSS), StarGAN-BlendFace-Stack (SBS), and SimSwap [222] for face swapping. For face reenactment, the techniques used are Talking Head Video [258], ATVG-Net [191], Motion-cos [259], and FOMM [105]. For face alteration, the methods include StyleGAN2 [11], MaskGAN [189], StarGAN2 [190], SC-FEGAN [192], and DiscoFaceGAN [260]. The Text2Image techniques involve Openjourney (OJ) [261], Stable Diffusion 1 (SD1), Stable Diffusion 2 (SD2), Stable Diffusion XL (SDXL) [262], Wenxin [263], Midjourney [264], DF-GAN [265], DALL·E, and DALL·E 3 [266]. For Text2Video, the techniques are AnimateDiff [267], AnimateLCM [268], Hotshot [269], Zeroscope [270], and Magic-

Time [271]. The Image2Image subset includes InstructPix2Pix (Pix2Pix), Stable Diffusion XL Refiner (SDXLR), and Stable Diffusion Image Variation (VD) [262]. The quality of the forgery samples is high, and the dataset size is 491.1 GB.

2.66. DF40 [172]

This dataset is composed of over 100,000 fake images and more than 1 million fake videos. It used 52,590 real videos (Table 10 in [172]). Four types of deepfakes were targeted, i.e., face swapping, face reenactment, entire face synthesis, and face editing (i.e., attribute manipulation). The quality of deepfakes is high. This dataset contains a diverse range of deepfake samples, which were generated using 40 different deepfake techniques. Face swapping methods include FSGAN [180], FaceSwap [183], SimSwap [222], InSwapper [272], BlendFace [257], UniFace [273], MobileSwap [19], E4S (editing for swapping) [218], FaceDancer [223], and DeepFaceLab [177]. Face reenactment techniques comprise FOMM [105], FS_vid2vid (few shot video-to-video) [274], Wav2Lip [187], MRAA (motion representations for articulated animation) [275], OneShot [276], PIRender (portrait image neural renderer) [277], TPSMM (thin-plate spline motion model) [278], LIA (latent image animator) [279], DaGAN (depth-aware GAN) [280], SadTalker (stylized audio-driven talking-head) [281], MCNet (memory compensation network) [282], HyperReenact [283], and HeyGen [284]. Face synthesis techniques include VQGAN (vector-quantized GAN) [285], StyleGAN2 [11], StyleGAN3 [286], StyleGAN-XL (StyleGAN large-scale) [287], SD-2.1 (Stable-Diffusion-2.1) [288], DDPM (denoising diffusion probabilistic model) [289], RDDM (residual denoising diffusion model) [290], PixArt-α (transformer-based text-to-image diffusion model) [291], DiT-XL/2 (diffusion transformers large) [292], SiT-XL/2 (self-supervised vision transformer) [293], MidJounery6 [264], and WhichisReal [80]. Face editing methods encompass CollabDiff (collaborative diffusion) [294], e4e (encoder for editing) [295], StarGAN [178], StarGANv2 [190], and StyleCLIP (styleGAN contrastive language-image pre-training) [26]). Deepfake samples were generated using original/real samples from the FaceForensics++ [76], Celeb-DF [68], UADFV [64], VFHQ (high-quality video face dataset) [296], FFHQ [78], CelebA [297], FaceShifter [298], and Deeperforensics-1.0 [81] datasets.

3. Audio Deepfake Datasets

A variety of audio deepfake databases have been created to advance both the audio deepfake detection and technology of audio deepfakes. In this section, we provide a meticulous summary of existing audio deepfake databases. We extensively reviewed multiple research articles and database repositories to provide comprehensive information, encompassing insights not commonly present in prior works. We have included all publicly reported and/or accessible datasets from 2018 to 2024, showcasing the advancements in the field of image and video deepfake datasets. A comparative analysis of existing audio deepfake datasets is depicted in Tables 22–27.

Tables 22–27 list the audio deepfake datasets in ascending order by year (first column) and then in alphabetical order by name within each year (second column) to provide a view of how the audio deepfake datasets and the audio deepfake field have progressed over time. This chronological arrangement aims to help readers observe the evolution of audio deepfake datasets, including deepfake types, languages of samples, durations, formats, quantities, quality, and the methods/tools used to generate deepfakes.

Table 22. Comparison and key characteristics of existing audio deepfake datasets. TTS = Text to Speech; VC = Voice Conversion; SS = Speech Synthesis.

Year	Dataset	No. of Real Samples	No. of Fake Samples	Total No. of Samples	Types of Deepfakes	Length	Language	Speaker Genders	Format	Methods Used to Generate Deepfake	Size of Dataset	Link to Download the Dataset
2018	Baidu Silicon Valley AI Lab cloned audio [12,299]	10 [12]	120 [12]	130 (C)	Voice cloning [12,299]	6 h [12]	English [12,299]	-	MP3 [12]	Encoder-based Neural Networks [299]	∼75.27 MB (C) [300]	https://audiodemos.github.io/ (accessed on 4 January 2024)
2018	M-AILABS [301]	-	German: 92,088, English (UK): 23,559, French: 82,813, Polish: 26,322, Ukrainian: 35,410, Russian: 20,495, Spanish: 45,530, English (US): 46,294, Italian: 26,519 (C) [301]	399,030 (C)	Speech Synthesis [301], TTS [302]	999.32 h [301]	German, English, Spanish, Italian, Ukrainian, Russian, French, Polish [301]	Male: 46.2% and Female: 53.8% (C)	WAV [301]	-	∼110.1 GB [301]	https://www.caito.de/2019/01/03/the-m-ailabs-speech-dataset/ (accessed on 4 January 2024)
2019	ASV Spoof 2019 [303]	39,000 [17]	15,600 [17]	55,200 [17]	VC and SS [303]	-	English [17]	Male: 43%, Female: 57%; 107 speakers [17,303]	FLAC [304]	Merlin, CURRENT, MaryTTS [303], LSTM, WaveRNN, WaveNet, WaveCycleGAN2 [16]	23.55 GB [304]	https://datashare.ed.ac.uk/handle/10283/3336 (accessed on 4 January 2024)
2019	Cloud2019 [305,306]	-	-	11,785 [307]	TTS [306]	-	English (C)	-	WAV, PCM [305]	Amazon AWS Polly, Google Cloud Standard, Google Cloud Wave Net, Microsoft Azure, IBM Watson [306,307]	-	-
2019	FoR [308]	111,000+ [17,308]	87,000+ [17,308]	198,000 + [17,308]	Speech Synthesis [308] and TTS [14,308]	150.3 h [14]	English [17, 308]	140 real and 33 fake speakers [17,308]	WAV [14,308]	Deep Voice 3, Amazon AWS Polly, Baidu TTS, Google, Traditional TTS, Google Cloud TTS, Google Wavenet TTS, Microsoft Azure TTS [308]	16.1 GB [309]	https://bil.eecs.yorku.ca/datasets/ (accessed on 4 January 2024)

Table 23. Comparison and key characteristics of existing audio deepfake datasets. TTS = Text to Speech; VC = Voice Conversion; SS = Speech Synthesis.

Year	Dataset	No. of Real Samples	No. of Fake Samples	Total No. of Samples	Types of Deepfakes	Length	Language	Speaker Genders	Format	Methods Used to Generate Deepfake	Size of Dataset	Link to Download the Dataset
2019	H-Voice [310]	Imitation: 3332, Synthetic: 4 [299]	Imitation: 3264, Synthetic: 72 [299]	6672 [299,310]	TTS [299]	-	Spanish, English, Portuguese, French, Tagalog [299]	-	WAV, ONG [310]	Imitation method, Deep Voice [310]	370 MB [300,311–313]	Link 1: https://data.mendeley.com/datasets/ytkv9w92t6/1; Link 2: https://www.kaggle.com/datasets/dduongtrandai/hvoice-fake-voice; Link 3: https://www.sciencedirect.com/science/article/pii/S2352340920302250?via%3Dihub (accessed on 4 January 2024)
2020	Ar-DAD [314]	15,810 [310,314]	379 [310,314]	16,209 [310]	Imitation [310,314]	-	Arabic [310,314]	30 speakers [310,314]	WAV [310,314]	-	9.37 GB [315]	https://data.mendeley.com/datasets/3kndp5vs6b/1 (accessed on 4 January 2024)
2020	VCC 2020 [316]	2660 [317]	6120 [317]	8780 (C)	Voice Conversion [316,317]	-	English, Finnish, German, Mandarin [316]	Male: 50%, Female: 50%, 56 speakers: 7 male and 7 female in each language [316]	wav [318]	Neural vocoders, encoder–decoder, GANs, seq2seq networks [316]	195.5 MB [318]	https://zenodo.org/record/4345689 (accessed on 4 January 2024)
2021	ASVspoof 2021 [319,320]	1,383,820 [17]	130,032 [17]	1,513,852 [17]	Speech Synthesis, Voice Conversion [320]	-	English [17]	Male: 41.5%, Female: 58.5%; 48 speaker [320]	MP3, M4A, OGG [320]	Vocoders [320]	34.5 GB [321–323]	https://zenodo.org/records/4835108; https://zenodo.org/records/4834716 (accessed on 27 April 2024)

Table 24. Comparison and key characteristics of existing audio deepfake datasets. TTS = Text to Speech; VC = Voice Conversion; SS = Speech Synthesis.

Year	Dataset	No. of Real Samples	No. of Fake Samples	Total No. of Samples	Types of Deepfakes	Length	Language	Speaker Genders	Format	Methods Used to Generate Deepfake	Size of Dataset	Link to Download the Dataset
2021	FMFCC-A [324]	10,000 [324]	40,000 [324]	50,000 [324]	TTS, VC [324]	-	Mandarian [324]	Female: 55.73%, Male: 44.27%; 131 speakers [324]	WAV, MP3, AAC [324]	Alibaba TTS, Biaobei TTS, Blackberry TTS, FastSpeech, Iflytek TTS, IBM Waston TTS, Lingyun TTS, Tacotron, Baidu TTS, Sibichi TTS, GAN-TTS [324]	3.31 GB [325]	https://ieee-dataport.org/documents/fmfcc-dataset (accessed on 4 January 2024)
2021	WaveFake [326]	-	117,985 [326]	117,985 [326]	TTS [326]	∼196 h [326]	English, Japanese [17,326]	2 [14,326]	WAV [326]	MelGAN, Parallel WaveGAN, Multi-band MelGAN, Full-band MelGAN, HiFi-GAN, WaveG-low [326]	28.9 GB [327]	https://zenodo.org/records/5642694 (accessed on 4 January 2024)
2022	ADD 2022 [328]	LF: 36,953, PF: 23,897, FG-D: 46,871 [14]	LF: 123,932, PF: 127,414, FG-D: 243,537 [14]	LF: 160,885, PF: 151,311, FG-D: 290,408 (C)	Partial fake, TTS, VC [14,328]	LF: 222 h, PF: 201.8 h, FG-D: 396 h [14]	Chinese [328]	PF: >200 real speakers and >200 fake speakers; LF and FG-D: >300 fake and >400 real speakers [14]	WAV [14]	-	3.73 GB [329]	http://addchallenge.cn/download (accessed on 5 January 2024)
2022	CFAD [330]	115,800 [330]	Noisy: 115,800; Codec: 115,800 [330]	347,400 [330]	TTS, partial fake [330]	-	Chinese [330]	Real: 1212; Noisy: 1212, Codec: 1212 speakers [330]	MP3, M4A, OGG, FLAC, AAC, WMA [330]	STRAIGHT, Griffin-Lim, LPCNet, WaveNet, neural-vocoder-based system (PWG, HifiGAN, Multiband-MelGAN, Style-MelGAN), WORLD, FastSpeech-HifiGAN, Tacotron-HifiGAN, Partially Fake [330]	29.9 GB [331]	https://zenodo.org/record/6635521 (accessed on 5 January 2024)

Table 25. Comparison and key characteristics of existing audio deepfake datasets. TTS = Text to Speech; VC = Voice Conversion; SS = Speech Synthesis.

Year	Dataset	No. of Real Samples	No. of Fake Samples	Total No. of Samples	Types of Deepfakes	Length	Language	Speaker Genders	Format	Methods Used to Generate Deepfake	Size of Dataset	Link to Download the Dataset
2022	FakeAV Celeb [116]	500 [116]	19,500 [116]	20,000 [116]	TTS [116]	-	English [116]	Male: 50%, Female: 50%; 500 speakers [116]	MP4 [116]	SV2TTS [116]	-	https://sites.google.com/view/fakeavcelebdash-lab/ (accessed on 5 January 2024)
2022	In-the-wild [332]	19,963 [14]	11,816 [14]	31,779 (C)	TTS [14,332]	38 h (20.8 h of bona fide and 17.2 h of manipulated audio) [332]	English [332]	58 genuine and 58 fake speakers [14,332]	WAV [14,332]	19 distinct TTS synthesis algorithms [332]	7.6 GB [333]	https://deepfake-demo.aisec.fraunhofer.de/in_the_wild (accessed on 4 January 2024)
2022	Lav-DF [150]	36,431 [150]	99,873 [150]	136,304 [150]	TTS [150]	-	English [150]	153 speakers [150]	MP4 [150]	Content-driven RE, Tacotron 2, SV2TTS [150]	24 GB [151]	https://drive.google.com/file/d/1-OQ-NDtdEyqHNLaZU1Lt9Upk5wVqfYJw/view (accessed on 5 January 2024)
2022	TIMIT [334]	2342 [334]	-	2342 [334]	TTS [334]	-	English [334]	630 speakers [334]	WAV [335]	Google TTS, Tacotron-2, MelGAN [334]	3 GB [335]	https://conradsanderson.id.au/vidtimit/ , https://zenodo.org/record/158963 (accessed on 4 January 2024)
2023	ADD 2023 [336]	LR: 46,554 AR: 14,907 FG-D: 172,819 [14]	LR: 65,449 AR: 95,383 FG-D: 113,042 [14]	LR:112,003 AR:110,290 FG-D:285,861 (C)	TTS, VC, Partial fake [14,336]	LR: 131.2 h; AR: 194.5 h; FG-D: 394.7 [14]	Chinese [14]	LR: >200 real and >200 fake speakers, AR: >500 real and >500 fake speakers, FG-D: >1000 real speakers and >500 fake speakers [14]	WAV [14]	-	-	http://addchallenge.cn/add2023 (accessed on 5 January 2024)
2023	AV-Deepfake 1M [65]	286,721 [65]	860,039 [65]	1,146,760 [65]	Voice cloning, Transcript manipulation [65]	1886 h [65]	English [156]	2068 Speakers [65]	MP4 [156]	VITS, YourTTS [65]	∼400 GB	https://github.com/ControlNet/AV-Deepfake1M (accessed on 1 January 2024)

Table 26. Comparison and key characteristics of existing audio deepfake datasets. TTS = Text to Speech; VC = Voice Conversion; SS = Speech Synthesis.

Year	Dataset	No. of Real Samples	No. of Fake Samples	Total No. of Samples	Types of Deepfakes	Length	Language	Speaker Genders	Format	Methods Used to Generate Deepfake	Size of Dataset	Link to Download the Dataset
2023	EmoFake [37]	17,500 [37]	39,900 [37]	57,400 [37]	EVC (Emotional Voice Conversion) [37]	-	English [37]	10 real and 10 fake speakers [37]	-	VAW-GAN-CWT, DeepEST, Seq2Seq-EVC, CycleGAN-EVC, CycleTrans-GAN, EmoCycleGAN, StarGAN-EVC [37]	-	https://drive.google.com/file/d/1aYPNVCVIBs6c9erbhT3U8YzClFEDsR/viewusp=sharing accessed on 5 January 2024)
2023	FSD [337]	200 [337]	450 [337]	650 [337]	VC, SS [337]	26.26 h [337]	Chinese [337]	Male: 50%, Female: 50%, Speakers: 27 male and 27 female [337]	WAV [338]	SO-VITS, NSF-HifiGAN with Snake, Shallow diffusion, DiffSinger, RVC [337]	-	https://github.com/xieyuankun/FSD-Dataset (accessed on 15 June 2024)
2023	Half-Truth [40]	53,612 [40]	53,612 (Fully Fake); 53,612 (Partially Fake) [40]	160,836 (C)	TTS [40]	-	Chinese [40]	Male: 80.28%, Female: 19.72%; 218 speakers [40]	WAV [339]	Global style token (GST) Tacotron [40]	8.1 GB [339]	https://zenodo.org/records/10377492 (accessed on 7 January 2024)
2023	SFR [340]	22,400 [340]	159,364 (C)	181,764 [340]	TTS [340]	526.53 h [340]	Chinese [340]	745 speakers [340]	MP3, FLAC, MP4, WMA, AMR, AAC, AVI, WMV, WAV, MOV, FLV [340]	Aispeech, Sogou, Alibaba Cloud, Baidu Ai Cloud, Databaker, Tencent Cloud, iFLYTEK [340]	-	-
2023	TIMIT-TTS [341]	-	80,000 [341]	80,000 [341]	TTS [341]	-	English [341]	Male: 50%, Female: 50%; 37 speakers [341]	WAV [341]	MelGAN, WaveRNN, Tacotron, Tacotron2, GlowTTS , FastSpeech2, FastPitch, TalkNet, MixerTTS, MixerTTS-X, VITS, SpeedySpeech, gTTS, Silero [341]	7.2 GB [342]	https://zenodo.org/records/6560159 (accessed on 5 January 2024)

Table 27. Comparison and key characteristics of existing audio deepfake datasets. TTS = Text to Speech; VC = Voice Conversion; SS = Speech Synthesis.

Year	Dataset	No. of Real Samples	No. of Fake Samples	Total No. of Samples	Types of Deepfakes	Length	Language	Speaker Genders	Format	Methods Used to Generate Deepfake	Size of Dataset	Link to Download the Dataset
2024	CD-ADD [343]	25,111 [343]	120,459 [343]	145,570 [343]	TTS [343]	>300 h [343]	English, Chinese [343]	-	-	VALL-E, YourTTS, WhisperSpeech, Seamless Expressive, Open-Voice	-	-
2024	Codecfake [344]	132,277 [344]	925,939 [344]	1,058,216 [344]	TTS [344]	-	English and Chinese [344]	-	WAV [345]	SoundStream, SpeechTokenizer, FunCodec, EnCodec, AudioDec, Academic-Codec, Descript-audio-codec (DAC) [344]	32.61 GB [345]	https://github.com/xieyuankun/Codecfake (accessed on 16 June 2024)
2024	CtrSVDD [346]	32,312 [346]	188,486 [346]	220,798 [346]	VC, SS [346]	307.98 h [346]	Mandarin and Japanese [346]	164 singers [346]	-	XiaoiceSing, VISinger, VISinger2, NNSVS, Naive RNN, Diff-Singer, ACESinger, NU, WavLM, ContentVec, MR-HuBERT, WavLabLM, Chinese HuBERT [346]	-	https://github.com/SVDDChallenge/CtrSVDD2024_Baseline (accessed on 15 June 2024)
2024	FakeSound [347]	39,597 [348]	3798 [347]	43,395 (C)	AM [347]	-	English [348, 349]	-	WAV [350]	AudioLDM1/AudioLDM2 + AudioSR [347]	986.6 MB [350]	https://github.com/FakeSoundData/FakeSound (accessed on 20 June 2024)
2024	SceneFake [351]	19,838 [351]	64,642 [351]	84,480 [351]	AM [351]	64 h [351]	English [351]	-	-	SSub, MMSE, Wiener, FullSubNet [351]	-	-
2024	SingFake [352]	15,488 [352]	11,466 [352]	26,954 [352]	VC, Voice Cloning [352]	58.33 h (28.93 h of bona fide and 29.40 h of deepfake) [352]	Mandarin, Cantonese, English, Spanish, Japanese, Persian [352]	40 speakers [352]	MP3, AAC, OPUS, VORBIS [352]	Wav2vec2 + AASIST [352]	-	https://github.com/yongyizang/SingFake/tree/main (accessed on 5 January 2024)

3.1. Baidu Silicon Valley AI Lab Cloned Audio [299]

This dataset was released by the Baidu Silicon Valley AI Lab. The set has 130 samples, including 10 real and 20 fake voice clips. The voice-cloning-based fake samples were created using encoder-based neural networks. The dataset includes 6 h of high-quality audio clips featuring multiple speakers. The samples are in MP3 format. The dataset is ∼75.27 MB in size.

3.2. M-AILABS [301]

The M-AILABS database spans an extensive duration of 999 h and 32 min. This compilation features a diverse array of languages, including German, Spanish, English, Italian, Russian, Ukrainian, French, and Polish. It includes voices from both male and female speakers. The samples are in WAV format. The total size of the dataset is approximately ∼110.1 GB.

3.3. ASV Spoof 2019 [303,353,354]

ASVspoof-2019 consists of two components, i.e., LA (logical access) and PA (physical access). Both components were derived from the VCTK base corpus [355], encompassing audio clips sourced from 107 speakers (61 females and 46 males). LA encompasses both voice conversion and speech synthesis samples, while in PA there is a combination of replay samples and genuine recordings. Both components are subdivided into three distinct sets (i.e., training, development, and evaluation). The fake samples were created using deep learning techniques like Merlin [356], CURRENT [357], MaryTTS [358], long short-term memory (LSTM) [359,360], WaveRNN [361], WaveNet [362], and WaveCycleGAN2 [363]. The samples are preserved in the FLAC format, and the dataset has a size of approximately 23.55 GB.

3.4. Cloud2019 [305,306]

This dataset has a total of 11,785 samples. These samples are derived from TTS (Text-to-Speech) cloud services, including Amazon AWS Polly (PO) [364], Google Cloud Standard (GS) [365], Google Cloud Wave Net (GW) [365], Microsoft Azure (AZ) [366], and IBM Watson (WA) [367]. The dataset primarily consists of English-language audio clips that were generated by automated systems using human voices as input. The audio files are in the WAV and PCM formats.

3.5. FoR (Fake or Real) [308]

The Fake or Real (FOR) dataset, introduced in 2019, is a significant collection in the domain of audio deepfake detection. This dataset comprises a total of 198,000+ samples, with 87,000+ being synthetic samples created through the utilization of Deep Voice 3 [368], Amazon AWS Polly, Baidu TTS, Google traditional TTS, Google cloud TTS, Microsoft Azure TTS, and Google Wavenet [365]. These samples collectively span a duration of 150.3 h and are in the English language. The dataset features voices with a distribution of 140 real and 33 fake speakers, and the samples are stored in WAV format. The database has a size of 16.1 GB.

3.6. H-Voice [310]

H-Voice is a dataset that includes a total of 6672 recordings, consisting of 3332 (in imitation section) and four (in synthetic section) original recordings, and 3264 (in imitation section) and 72 (in synthetic section) fake recordings. This dataset is designed for TTS applications and offers recordings in the WAV audio format. It covers a range of languages, including Spanish, English, Portuguese, French, and Tagalog, and features contributions from 84 different speakers. The voice recordings were generated using the imitation and deep voice methods. The dataset needs approximately 370 MB of storage space.

3.7. Ar-DAD (Arabic Diversified Audio Dataset) [314]

The 'Ar-DAD: Arabic Diversified Audio' dataset comprises a significant total of 16,209 audio recordings. This dataset primarily focuses on Arabic speech data and features contributions from 30 different speakers. In particular, the speakers comprise individuals from Arabic-speaking regions such as Saudi Arabia, Egypt, Kuwait, Yemen, UAE, and Sudan. The audio files are saved in the WAV format, making them versatile for various applications and research in Arabic language processing and speech synthesis. The recordings in this dataset were generated using the imitation method, and it takes approximately 9.37 GB of storage space.

3.8. VCC 2020 [316]

The VCC 2020 dataset is part of the Voice Conversion Challenge 2020 and encompasses multiple languages, including English, Finnish, German, and Mandarin, making it suitable for multilingual studies. The dataset is balanced in terms of gender, with a 50% representation of male and female voices. The voice-conversion-based deepfake samples were created via sequence-to-sequence (seq2seq) mapping networks, neural vocoders, GANs, and encoder–decoder networks. The audio files in this dataset are stored in the WAV format. The storage size of the dataset is approximately 195.5 MB.

3.9. ASVspoof 2021 [319,320]

The ASVspoof 2021dataset is a comprehensive and diverse collection of 1,513,852 audio samples in English language. There are 130,032 VC (Voice Conversion)- and TTS (Text-to-Speech)-based fraudulent samples that were created by employing VC attack algorithms [353] and Vocoders. It exhibits a balanced distribution of gender, with 41.5% male voices and 58.5% female voices. The dataset's audio files are stored in MP3, A4A, and OGG formats and have a storage size of approximately 34.5 GB. The logical access portion of the database can be accessed on [323].

3.10. FMFCC-A [324]

The FMFCC-A database was introduced in 2021 and comprises a total of 50,000 samples, of which 10,000 are real and 30,000 are fake. The dataset features the Mandarin language and the samples are available in WAV, AAC, and MP3 formats. For the generation of deepfakes, methods such as Text To Speech (TTS) (e.g., Fastspeech 2 [369]) and Voice Conversion (e.g., Ada-In VC [370]) were employed. Other systems used were Alibaba TTS, BlackBerry TTS, IBM Waston TTS, Tacotron [235], and GAN TTS [371]. The dataset occupies about 3.31 GB.

3.11. WaveFake [326]

This database comprises 117,985 synthesized audio clips. The samples encompass both English and Japanese languages. The TTS synthetic samples were produced by leveraging the LJSpeech [326] and Japanese Speech Corpus (JSUT) [372] databases, employing MelGAN [373], Parallel Wave-GAN (PWG) [374], Multiband MelGAN (MB-MelGAN) [375], Fullband MelGAN (FB-MelGAN), HiFi-GAN [376], and Wave-Glow [377] models. The synthetic samples, though high-quality and realistic, lack diversity as each sample has only one speaker. The samples are in the WAV format, and the total dataset size is approximately 28.9 GB.

3.12. ADD 2022 [328]

The ADD 2022 dataset constitutes a component of the ADD 2022 challenge. This database includes partially fake audio detection (PF) and low-quality fake audio detection (LF). LF comprises 36,953 genuine voices and 123,932 fabricated spoken words with real-world noises. In contrast, PF encompasses 127,414 audio samples that are manipulated or altered. The ADD dataset, which is openly accessible, is exclusively in the Chinese language. The dataset requires 3.73 GB of storage space. The findings from ADD 2022

indicate that a single method or model is not capable of addressing all types of fakes and generalization.

3.13. CFAD (Chinese Fake Audio Detection) [330]

Within the CFAD dataset, there are 347,400 samples contributed by a total of 1212 real and 1212 fake speakers (Table II in [330]). The TTS and partial fake audio clips in Chinese were generated using 12 vocoder techniques (i.e., STRAIGHT [378], Griffin-Lim [379], LPCNet [380], WaveNet [362], neural-vocoder-based system (PWG [374], HifiGAN [376], Multiband-MelGAN [375], Style-MelGAN [381]), WORLD [382], FastSpeech-HifiGAN [369], Tacotron-HifiGAN [235], and Partially Fake [40,383,384]). The audio files are available in multiple formats, including MP3, M4A, OGG, FLAC, AAC, and WMA. The dataset's storage size is approximately 29.9 GB.

3.14. FakeAVCeleb [116]

This database encompasses a collection of 20,000 audio samples with 500 real and 9500 fake samples in English language. The manipulated audio samples were generated by a real-time voice cloning method known as SV2TTS [20]. The dataset exhibits an equal distribution of male and female voices, with each gender contributing 50% of the dataset. The samples are available in MP4 format.

3.15. In-the-Wild [332]

This database is made up of total 31,779 audio samples, including 11,816 synthesized samples using 19 distinct TTS synthesis algorithms. There are 58 genuine and 58 fake English-speaking celebrities and political speakers. The total audio duration is 38 h. The samples are stored in WAV format, and the dataset size is approximately 7.6 GB. This study also used RawNet2 [385] and RawGAT-ST [386] techniques for audio deepfake detection.

3.16. Lav-DF (Localized Audio Visual DeepFake) [150]

This dataset comprises a total of 136,304 English audio clips, with 99,873 clips being fake segments. The dataset comprises 153 speakers. This dataset generated Text-to-Speech (TTS) deepfake content, employing content-driven Recurrent Encoder (RE), Tacotron 2 [387], and SV2TTS [20] techniques. The dataset contains audio files in MP4 format, requiring approximately 24 GB of storage.

3.17. TIMIT [334]

The TIMIT dataset [388] was expanded to create this dataset with fake audio samples. It features English speech data and includes recordings from 630 different speakers. The fake samples mimic human voice and were created by state-of-the-art neural speech synthesis techniques such as Google TTS [365], Tacotron-2 [387], and MelGAN [373]. The audio files are stored in WAV format, and the dataset size is 3 GB.

3.18. ADD 2023 [336]

The ADD 2023 dataset is an integral part of the ADD 2023 challenge. Audio Fake Game (FG), Manipulation Region Location (RL), and Deepfake Algorithm Recognition (AR) are three sub-challenges in the ADD 2023 challenge. In contrast to the ADD 2022 challenge, ADD 2023 shifts its focus from binary fake or real classification. Instead, it concentrated on the more intricate task of localizing manipulated intervals within a partially fake utterance and identifying the specific source responsible for generating any fake audio. This nuanced approach adds a layer of complexity to the evaluation and fosters advancements in understanding and addressing more sophisticated manipulations in audio data.

3.19. AV-Deepfake 1M [65]

This dataset is composed of 286,721 legitimate samples and 860,039 manipulated audio samples. The fakes are of voice cloning and manipulation of transcripts deepfake types, and they were generated using VITS (Variational Inference with adversarial learning for end-to-end Text-to-Speech) [215] and YourTTS [217]. This is an extensive audio dataset of 1886 h from 2068 speakers. The samples in this dataset are in MP4 format with a total size of ∼400 GB.

3.20. EmoFake [37]

This dataset was created using the English portion of the Emotional Speech Database dataset [389]. The EmoFake dataset consists of both emotion-manipulated audio and authentic emotional speech. In the EmoFake dataset, there are 57,400 samples in total, comprising 17,500 authentic recordings and 39,900 artificially generated instances. The fake samples were generated by transitioning the emotional state from a source emotion to a target emotion. The fake audios were produced using seven open-source Emotional Voice Conversion (EVC) models (i.e., VAW-GAN-CWT [390], Seq2Seq-EVC [391], CycleTransGAN [392], DeepEST [393], EmoCycleGAN [394], StarGAN-EVC [395], and CycleGAN-EVC [396]).

3.21. Fake Song Detection (FSD) [337]

This Chinese fake song detection dataset consists of 200 real and 450 fake songs. The fake songs were generated using five state-of-the-art singing voice conversion and singing voice synthesis techniques. The fake songs were produced employing SO-VITS [397], NSF-HifiGAN with Snake [398], SO-VITS with shallow diffusion [399], DiffSinger [399], and Retrieval-based Voice Conversion (RVC) [229]. The dataset features voices with a distribution of 27 male and 27 female singers, and the samples are in WAV format.

3.22. Half-Truth [40]

This dataset was meticulously curated to encourage research in identifying partially fake audio utterances, where any single sample contains elements of both truth and falsehood. The partially fake samples were generated by splicing fractions of synthetic audio into the original speech. The fake audio samples were produced utilizing the style token (GST) Tacotron technique [383,384]. This dataset is in Chinese with 175 female and 43 male speakers. The samples are conveniently provided in WAV format, and the total size of the dataset is 8.1 GB.

3.23. SFR (System Fingerprint Recognition) [340]

Established in 2023, this database comprises 181,764 audio clips in total. These recordings span 526.53 h of Chinese speech data and are distributed across 745 different speakers. There are 22,400 authentic samples and 159,364 fabricated samples. The fake samples were generated through the seven TTS systems, i.e., Aispeech [400], Sogou [401], Alibaba Cloud [402], Baidu Ai Cloud [403], Databaker [404], Tencent Cloud [405], and iFLYTEK [406]. All audio files are stored in the MP3, MP4, WMA, AMR, AAC, AVI, WMV, MOV, and FLV formats.

3.24. TIMIT-TTS [341]

This database employed the VidTIMIT [407] and DeepfakeTIMIT [51] datasets to produce TTS deepfakes, comprising a total of 80,000 synthetic audio tracks. Twelve distinct state-of-the-art TTS methods (MelGAN [373], WaveRNN [361], Tacotron [235], Tacotron 2 [387], GlowTTS [408], FastSpeech2 [369], FastPitch [409], TalkNet [410], MixerTTS [411], MixerTTS-X [411], VITS [215], SpeedySpeech [412], gTTS [413], and Silero [414]) were utilized to generate the TTS deepfakes. To elevate the authenticity of the generated tracks, a Dynamic Time Warping (DTW) [415] step was incorporated. The database features a balanced representation of male (50%) and female (50%) voices. The audio files conform to the WAV format, and the database has a total size of approximately 7.2 GB.

3.25. Cross-Domain Audio Deepfake Detection (CD-ADD) [343]

The CD-ADD dataset encompasses 25,111 real and 120,459 audio deepfake samples, totaling 145,570 samples. The dataset has more than 300 h of speech samples generated by five zero-shot TTS models that are grouped in two types: decoder-only (i.e., VALL-E [236]) and encoder–decoder (i.e., YourTTS [217], WhisperSpeech [416], Seamless Expressive [417], and Open-Voice [418]).

3.26. Codecfake [344]

The Codecfake dataset was created to detect audio language model-based deepfake audio. It encompasses over 1 M audio samples in both English and Chinese languages. Specifically, there are 1,058,216 audio samples, with 132,277 and 925,939 being real and fake, respectively. The real audio samples were sourced from the VCTK [355] and AISHELL3 [419] databases. The fake samples were fabricated using seven different codec techniques, i.e., SoundStream [420], SpeechTokenizer [421], FunCodec [422], EnCodec [423], AudioDec [424], AcademicCodec [425], and Descript-audio-codec (DAC) [426]. The audio samples are in WAV format, and the total size of the dataset is 32.61 GB.

3.27. Controlled Singing Voice Deepfake Detection (CtrSVDD) [346]

The CtrSVDD is a collection of diverse bona fide and deepfake singing vocals. For 164 singers, this dataset has 188,486 deepfake song samples and 32,312 bona fide song clips for 164 singers. The total number of samples is 220,798 mono vocal clips totaling 307.98 h. The bona fide singing vocals were taken from mandarin singing datasets (i.e., Opencpop [427], M4Singer [428], Kising [429], and official ACE-Studio release [430]) and Japanese singing datasets (i.e., Ofuton-P [431], Oniku Kurumi [432], Kiritan [433], and JVS-MuSiC [434]). The deepfake vocals were generated using both Singing voice synthesis (SVS) and Singing voice conversion (SVC) methods. The SVS systems used are XiaoiceSing [435], VISinger [436], VISinger2 [437], neural network (NN)-based SVS (NNSVS) [438], Naive RNN [439], DiffSinger [399], and ACESinger [429,430]. The SVC systems utilized are Nagoya University (NU) [440], WavLM [441], ContentVec [442], MR-HuBERT [443], WavLabLM [444], and Chinese HuBERT [445].

3.28. FakeSound [347]

The FakeSound dataset consists of 39,597 genuine audio samples and 3798 manipulated audio samples. To create deepfake samples, a grounding model identifies and masks key areas in legitimate audio based on caption information. The generation model then recreates these key regions, substituting them to generate convincing, realistic deepfake audio. In particular, open-source models (i.e., AudioLDM1 [446] and AudioLDM2 [447]) inpaint masked portions using input text and remaining audio. AudioSR [448] upscales for realism and quality; then, the regenerated segments are combined with the original audio. The audio samples are in WAV format, totaling 986.6 MB in size.

3.29. SceneFake [351]

The SceneFake dataset was created to detect scene-manipulated utterances. This database is assembled from manipulated audios, which were generated by only tampering with the acoustic scene of real utterances by employing speech enhancement technologies. It includes both fake and genuine utterances involving various scenes. The dataset was developed using the logical access (LA) section of ASVspoof 2019 and the acoustic scene dataset from the DCASE 2022 challenge. It contains 19,838 real samples, 64,642 fake samples, and 84,480 total samples, all in English. The dataset features six types of acoustic scenes, i.e., Airport, Bus, Park, Public, Shopping, and Station. The fake utterances were manipulated by utilizing four kinds of speech enhancement methods, i.e., Spectral Subtraction (SSub) [449,450], Minimum Mean Square Error (MMSE) [449,450], Wiener filtering [449,450], and Full and band network (FullSubNet) [451].

3.30. SingFake [352]

The SingFake dataset, introduced in 2024, consists of a total of 26,954 audio clips, with 15,488 clips categorized as real and 11,466 clips categorized as fake. The dataset has a cumulative duration of 58.33 h and encompasses multiple languages, including Mandarin, Cantonese, English, Spanish, and Japanese. All audio files from the 40 speakers are stored in the MP3, AAC, OPUS, and VORBIS formats.

Evolution and Transition to Future Prospects of Deepfake Datasets: The deepfake datasets described in Sections 2 and 3 consist of extensive collections of visual and audio deepfake media, which have become crucial in advancing AI's capabilities to both create and detect deepfakes. This collection is an invaluable resource for those pursuing an in-depth and exhaustive analysis of the subject matter. Overall, during the preliminary years, these datasets were limited and less diverse, but they have progressively become more sophisticated and expansive. In spite of the remarkable progress, there is a growing demand for larger and more comprehensive databases that cover a broader spectrum of scenarios, media types, and characteristics to keep pace with advancing technology, as discussed in Sections 4.2.6, 4.10, 4.15 and 4.22.

Considerations and Strategies for Selecting Deepfake Datasets: Still, the crucial question is how one decides which databases to select from the available options for their research or study. The selection of deepfake datasets involves understanding the purpose of one's study, the types of deepfake techniques one aims to address, and the quality and diversity of the dataset needed for robust model development. To this aim, first, define your research goal (i.e., whether you focus on single deepfake detection or generalized multiple deepfakes detection). Next, determine which types of deepfakes you need (e.g., audio or visual deepfakes, multimodal deepfakes, or full body deepfakes). Look for datasets with different quality, diverse samples, and sizes (e.g., high-resolution images or videos and sophisticated audio files and formats with a variety of ethnicities, ages, and environments) to ensure robust and advanced model development. Larger and more diverse datasets can improve the generalizability of your model. Moreover, choose datasets that are commonly used in the field to facilitate benchmarking and comparison with other research. Benchmarking datasets (e.g., Deepfake Detection Challenge [85] and DeepFake Game Competition [108]) are widely recognized and used within the research community, which allows for standardized evaluation and comparison of different methods and models and facilitates progress and innovation in the field. All in all, to substantiate the efficacy, robustness, and benefits of the developed deepfake framework, it is crucial to report experimental analyses using heterogeneous datasets from various time periods. For instance, employing one dataset from the early years, another from the middle years, and a recent dataset from the last two years (as more advanced tools are typically used to generate deepfakes in recent datasets) would provide a comprehensive evaluation. This selection of diverse databases will objectively assess the dynamism and effectiveness of the developed algorithms under a cross-database setting, wherein training is performed on one or more databases and testing is conducted on different ones.

4. Open Issues and Potential Research Directions

Despite significant progress in the deepfake field, there remain various noteworthy concerns surrounding deepfake generation and detection frameworks [12,14,66,319,452–467]. This section highlights both key challenges and future research directions in the field of deepfake technology. This section roughly follows a topical structure, encompassing themes like benchmarking, next-generation deepfake detectors and generators, critical factors (e.g., source identification and fairness), emerging frontiers in deepfake technology, regulatory landscapes, delving into the complexities of audio deepfake techniques, combating deepfake vulnerabilities and cybersecurity threats, and the importance of cross-disciplinary insights for effective mitigation and awareness. Some content in subsections within these themes may seem to overlap in subject matter, but they are intentionally structured this way to emphasize their significance without diluting their focus and importance.

4.1. Assessing Effectiveness: Benchmarking and Metrics for Evaluation

The accuracy of audio and visual deepfake and manipulation methods grapples with fundamental issues, including the evaluation, configuration, and comparison of algorithms. A comprehensive evaluation framework is essential to assess and rate the capacity of deepfake generation and detection techniques. Practitioners and researchers could potentially accomplish it by (i) formulating protocols and tools specifically designed to estimate the quality of generated deepfakes, performance of deepfake detectors, vulnerability analysis of deepfake detectors against adversarial attacks, and robustness of speech and face recognition systems under deepfakes; (ii) creating standardized common criteria; and (iii) establishing an online and open platform for transparent and independent evaluations of systems against validated benchmarks. Professionals and researchers are encouraged to devise novel performance evaluation matrix-, method-, and security- and non-security-related error rates, as well as propose some unified frameworks, universal taxonomy, and shared vocabulary for deepfake systems. We need to design qualitative and qualitative evaluation metrics.

Protocols ought to establish a foundation for current and future deepfake generation and detection methods, addressing both existing (known) and new (unknown) attribute manipulations and attacks. This is crucial for continuous and timely progress in the field. Furthermore, it is imperative to equally prioritize evaluation metrics that differentiate false positives in deepfake detection and speech and face recognition systems because false positives in deepfake detection occur when genuine content is incorrectly labeled as fake. This can potentially lead to unwarranted scrutiny or filtering of legitimate content, thereby affecting the reliability of detection algorithms. Conversely, false positives in speech or face recognition systems occur when the system incorrectly identifies an unauthorized user as legitimate. This can lead to unauthorized access being granted, thereby compromising security, privacy, and data integrity. Also, protocols should integrate sociocultural parameters such as human perceptions, human detectors' background and knowledge, and privacy.

Studies [4,51] have shown that higher quality samples are not only difficult to detect but also notably degrade the performance of face recognition systems. Thus, the online and open platform should provide datasets with varying qualities as well as options to generate deepfakes of different qualities using freely available tools or commercial software that may be automatic or require human intervention. To report baseline accurately without conveying any misleading impression of progress, it is imperative to define common criteria to assess deepfake quality, computational complexity, decision-making, performance, and policies.

All in all, a few initial attempts (e.g., [468–471]) have been made in this direction. However, there is still a deficiency in conducting large-scale analyses of existing evaluation metrics in the field of deepfakes to determine their reasonability. Any such study should also consider human capabilities and disparities between machines and humans for face manipulation and deepfake. Streamlined benchmarking platforms are essential, incorporating modular design, extensibility, interpretability, vulnerability assessment, fairness evaluation, a diverse range of generators and detectors, and strong analytical capabilities.

4.2. Future Deepfake Detection Technologies

This subsection focuses on future cutting-edge research on frameworks to detect and combat increasingly sophisticated deepfake contents. Specifically, the generalization capability of deepfake countermeasures in Section 4.2.1, identity-aware deepfake detection in Section 4.2.3, next-generation detection methods in Section 4.2.4, acoustic constituent-aware audio deepfake detectors in Section 4.2.5, multimodal deepfake detection in Section 4.2.6, and modality-neutral deepfake detection in Section 4.2.7 are discussed.

4.2.1. Generalization Capability of Deepfake Countermeasures

Many current deepfake detection models encounter difficulties in accurately identifying deepfakes and digitally manipulated speech and faces in datasets that deviate from their

original training data. Namely, the performance of detectors sharply drops upon encountering novel deepfakes types, databases, manipulations, variations, attacks, and speech or face generator models that were not utilized during the training phase [87,110,472,473]. This is attributed to the fact that previous approaches have predominantly concentrated on particular artifacts, deepfake scenarios, closed-set settings, and supervised learning, making them susceptible to overfitting. Besides overfitting, some other factors that may contribute to poor generalization are model architecture and complexity, underfitting, feature selection and engineering, hyperparameter tuning, data preprocessing, and dataset quality and representativeness. Attackers can take advantage of the lack of generalization. Generalizability is crucial for staying ahead of deepfake generators in the continuous development of effective countermeasures. There has been limited research aimed at addressing the generalizability of deepfake countermeasures. However, it is evident that we have not achieved generalizable deepfake detection. The problem of generalization is exacerbated by unknown deepfakes, temporal dynamics, and adversarial examples' transferability. The rapid development of deepfake detectors with broad generalizability is the need of the hour, ensuring effectiveness across diverse deepfake datasets, types, manipulations, resolutions, audio/visual synthesizers, apps, and tools.

For better understanding and future solutions, studies should strive to address questions such as 'to what extent do current deepfake detectors expand their capabilities in identifying deepfakes across diverse datasets?' and 'do deepfake countermeasures inadvertently learn unwanted features that impede generalization?'. Also, exploring the synergies between deepfake detectors and augmentation methods that contribute to better generalizability, and asking why deepfake detectors struggle to generalize could provide valuable insights. To confront the challenge of generalization capability, we need more robust, scalable, and adaptable deepfake detectors. Future research endeavors may center around enhancing the generalizability of deepfake countermeasures through the implementation of contrastive learning, meta learning, ensemble learning, active learning, semi-supervised learning, unsupervised learning, continuous learning, one class learning, causality learning (i.e., enhancing contributing neurons of generalizability, e.g., partial neurons/layer freezing), interferential learning (i.e., carefully applying strategies to avoid acquiring unwanted features, e.g., guided DropConnect), hybrid learning, zero-shot learning, few shot learning, open set recognition, human-in-the-loop, multi-attention, cross-modality attention, non-identity representations learning (e.g., deep information decomposition), multi-source to multi-target learning, universal features learning, and hierarchical learning methodologies.

4.2.2. Identity-Aware Deepfake Detection

The landscape of deepfake detection primarily revolves around 'artifact-driven' techniques or unnaturalness generation schemes. Thereby, detectors function well within the close-set context (i.e., known artifacts, manipulations, and attributes) but work poorly in the open-set context (i.e., unknown artifacts, manipulations, and attributes). New deepfake generators will continue to grow, producing deepfakes with novel manipulations and artifacts. One promising and reliable solution is 'identity-driven' deepfake detection, which learns identity implicit features rather than specific artifacts [25,163,474,475]. However, 'Identity-driven' deepfake detection techniques also struggle to perform effectively in open-set scenes, particularly when the detectors encounter previously unseen identities. Audio and multimodal identity-driven deepfake detection schemes do not currently exist. To develop innovative identity-aware deepfake detection technology, various multi-feature frameworks can be explored. These frameworks may incorporate components such as 3D morphable models, temporal attributes analysis, identity embeddings, metric learning techniques, adversarial training methodologies, attention mechanisms, continual learning approaches, self-supervised learning methods, federated learning strategies, graph neural networks, and secure multi-party computation techniques. Some key use cases of identity-aware deepfake detection could be content moderation (e.g., social platforms using identity-aware deepfake detection to automatically filter and flag fake videos, swiftly

removing misinformation and malicious content), evidence authentication (e.g., verifying the video authenticity evidence utilized in court cases to ensure that deepfake videos are not used to mislead the judicial process), fact-checking (i.e., news organizations employing this technology to verify video authenticity before broadcasting and preventing the spread of fake news and manipulated content), and fraud prevention (e.g., banks and financial institutions can use it to verify the identities of individuals in video banking services to avoid identity fraud and account takeovers).

4.2.3. Detection Methods Scarcity for Non-English Languages

Current research significantly prioritizes the crafting of detection techniques aimed at identifying deepfakes in the English language, particularly those in the audio domain. There is an acute shortage of frameworks specifically crafted for other major languages, such as Arabic, Hindi, Chinese, and Spanish, with estimated native speaker populations of 362 million, 344 million, 1.3 billion, and 485 million, respectively. This highlights a significant gap in existing research. The idiosyncrasies of alphabet pronunciation in each language present challenges that traditional English audio/video/multimodal deepfake detectors and audio/video/multimodal processing are not adept at handling. Also, non-English (audio/video/multimodal) deepfake databases are notably scarce. The research community should bridge this important gap by introducing innovative (audio/video/multimodal) deepfake detection methods capable of identifying languages beyond English. For instance, such novel, robust, and enhanced deepfake detection methods for Spanish, Arabic, Hindi, and Chinese could be integrated into social media platforms, empowering users to discern manipulated content, thereby increasing trust and reliability in digital communication.

4.2.4. Next-Generation Detection Methods

The rivalry between deepfake detection and generation resembles a dynamic cat-and-mouse game. Current static methods are not reliable, flexible, sustainable, and robust against attacks. It is crucial to devise more advanced detection schemes to overcome the limitations in prior audio/face-forensic technologies. The next-generation deepfake and audio/face manipulation detection algorithms should possess the capability to tackle diverse audio/facial features, sizes, and resolutions. They should be trainable with small databases without a loss of efficacy, be resilience to adversarial examples/attacks, function in real-time, and support multi-modal inputs and outputs. To achieve this goal, ensemble learning, multitask learning, adaptive schemes, multi-feature and multi-level fusion, as well as consideration of local and global attributes hold significant potential.

4.2.5. Acoustic Constituent-Aware Audio Deepfake Detectors

In the literature, a common observation is that most audio deepfake detection studies overlooked significant factors (e.g., accent, dialect, speech rate, pitch, tone, and emotional state), which could adversely affect the detection accuracy. Studies reveal the negative influence of such factors on speaker recognition performance [476]—there is a lack of sufficient research on this subject in audio deepfakes so far. More studies are warranted to understand the intricate effects of acoustic constituents like dialect, tone, accent, pitch, speech rate, emotional state, etc., both individually and in combination, on the efficacy of audio deepfake detectors. To assess its impact, one can evaluate the detection accuracy across varying levels of acoustic constituent sophistication, real-world applicability under diverse audio qualities and sources, computational speed and efficiency, comparison with current solutions, scalability and adoption potential, and ethical considerations regarding potential misuse.

4.2.6. Multimodal Deepfake Detection

Unimodal deepfake detectors, specializing in either visual or audio fake detection, constitute the prevailing majority within the current landscape of deepfake detection. As a result, they exhibit the poorest performance for multimodal deepfakes. Multimodal deep-

fakes are forged video samples that contain both fake visual and corresponding synthesized lip-synced fake audio. Although a handful of multimodal deepfake detectors have been proposed (e.g., [477,478]), their proficiency in identifying multimodal manipulations is notably substandard. The shortage of diverse and high-quality multimodal deepfake datasets restricts advancements in multimodal deepfake detection.

To catalyze increased efforts in the realm of multimodal deepfake detection studies, a greater number of diverse multimodal datasets should be produced and shared publicly for research purposes. These datasets should include types of forgeries (e.g., real audio paired with manipulated video, real video coupled with manipulated audio, manipulated video and manipulated audio, and manipulated partial portions) and a broad spectrum of event/context categories (e.g., indoor, outdoor, sports, and politics). Additionally, it is crucial to create public databases containing highly realistic multimodal deepfakes, wherein entire multimodal samples are meticulously crafted by AI. Generating cohesive and convincing multimodal deepfakes remains a challenge for contemporary techniques, particularly in ensuring the consistency and alignment of fake visuals and fake speech.

Future multimodal deepfake detectors may be grounded in diffusion-based multimodal features, multimodal capsule networks, multimodal graphs, multimodal variational autoencoders, multi-scale features, new audiovisual deepfake features, time-aware neural networks, novel disjoint and joint multimodal training, cross-model joint filters, multimodal multi-attention networks, and identity- and context-aware fusion mechanisms. Researchers should also develop ensemble-based multimodal deepfake detection systems to ensure superior generalization to unseen deepfakes and explainability. Additionally, exploring methods to improve the generalizability of multimodal deepfake detection methods to unseen deepfakes will be a key area of investigation.

4.2.7. Modality-Neutral Deepfake Detection

Deepfakes have evolved from singular modality manipulation to multimodal falsified contents, allowing for the fabrication of audio, visual, or combined audio–visual elements. Employing two separate unimodal detectors (i.e., one for audio and other for visual element) for audiovisual deepfakes either miss potential cross-modal forgery cues or introduce higher computational complexity [479,480]. Conventional multimodal deepfake detection methods often involve correlating audio and visual modalities, with a prerequisite for the concurrent existence of audio and visual modalities. In reality, deepfake samples may exhibit the absence of certain modalities, where either modality is missing or present at a time. Given the increasing prevalence of audiovisual deepfake content, it is essential to devise modality-agnostic detectors capable of efficiently thwarting multimodal deepfakes, regardless of the type or number of modalities at play. Future multimodal deepfake detection methods need to be cohesive, considering the intermodal relationships, and capable of addressing scenarios with missing modalities and the manipulation of any or both modalities. In these frameworks, detectors should assign two separate labels to accurately specify whether the audio, visual, or both modalities have undergone manipulation. Some potential directions for future research in this field include investigating multi-modal neural networks and ensemble frameworks that can integrate features from different sources to identify inconsistencies across modalities, as well as exploring domain adaptation and transfer learning methods that allow knowledge learned from detecting deepfakes in one modality (e.g., images) to be applied to another (e.g., audio).

4.3. Future Deepfake Generation Technologies

This subsection explores cutting-edge research on future frameworks for generating sophisticated deepfake content. In particular, next-generation audio/face manipulation and deepfake generators in Section 4.3.1, audio/facial attribute manipulation techniques in Section 4.3.2, and real-time deepfakes generation in Section 4.3.3 are covered.

4.3.1. Next-Generation Audio or Face Manipulation and Deepfake Generators

A dynamic and healthy interplay between deepfake generators and deepfake detectors may fuel great advances in designing next-generation multimedia forensic tools. The contemporary deepfake and audio/face manipulation methods are marked by several practical deficiencies, such as not being equipped for lifelike ultra-high-resolution sample creation, audio/face attribution manipulations constrained by the training dataset (i.e., new attributes could not be created if they are not part of training set), incompetent to manage scenarios involving occlusion, audiovisual continuity problems (i.e., jarring frame transitions and inconsistent physiological signals), are easily overfitted (e.g., training process exhibiting instability), poor generalization to unfamiliar situations, generated samples with artifacts or modal fingerprints being prone to easy detection by digital forensic tools, longer training time, higher generation (processing) time (especially for high-quality and -resolution samples), and requirement of sophisticated computing infrastructure.

Next-generation deepfake generation frameworks offer ample opportunities for improvement and innovation, for instance, developing automated and swift generators that leverage both deep learning and computer graphics techniques to create highly realistic samples. Similarly, they require generators that need a small training database; craft lifelike movies from few audios/photographs or from very short clips; maintain unbiasedness towards the dataset; efficiently handle occlusion, pose, resolutions, audio noises, audio distortions, and orientations, and class imbalances and demographic factors during and after training; and improve the synchronicity and temporal consistency of audiovisual deepfake content, particularly in the context of language-translated manipulation. Also, an imperative need exists for a computationally efficient speech/face generator designed to run seamlessly on edge devices or smartphones. Another avenue to explore is creating real-time 3D deepfakes with improved fidelity, head movements, and speech/facial expressions using advanced generators. Future 'entire sample synthesis' methods should not only generate audio or audiovisual samples without traces of identity-related indicators from the speeches/faces used in training (i.e., identity-agnostic schemes) but also be multi-domain and optimized for use on mobile devices.

Contrary to current deepfake generators, next-generation methods are poised to advance into full-body manipulations, incorporating changes in gestures, body pose, and more [211]. Also, there is a limited number of frameworks dedicated to producing obvious deepfakes and speech/face manipulations (i.e., human face with tusks) or creating target-specific individual audio or joint audio–visual deepfakes with a heightened level of realism and naturalness. Furthermore, designing dedicated hardware for deepfake generation and establishing fresh empirical metrics for efficacy offer exciting avenues for advancement.

Considering the viewpoint of the red team, future generators should possess the capability to generate samples devoid of subtle traceable information (i.e., fingerprints), speech/pixel inconsistencies, or unusual audio/texture artifacts, as these elements can potentially expose deepfakes to detection methods. Deepfake generators equipped with built-in mechanisms (e.g., traces tailored loss functions or layers) to mask or eliminate these elements without compromising quality can effectively circumvent proactive defenses. In essence, the competition between sophisticated generators and detectors will drive progress in both domains, particularly in developing robust detection frameworks and creating large-scale deepfake and face manipulation datasets.

4.3.2. Audio/Facial Attribute Manipulation Techniques

Audio/facial attribute schemes can be grouped into two main types: audio/facial attribute estimation (A/FAE) and audio/facial attribute manipulation (A/FAM). A/FAE is utilized to identify the presence of specific audio or visual attributes within a sample, while A/FAM involves modifying speeches or faces by synthesizing or removing attributes using generative models [453]. Existing audio/facial attribute techniques fail to successfully deal with audio, image, or video sizes, masks, vocal peripherals, obscurity, vocal cavity,

resolutions, and occlusions in the wild. The A/FAM methodologies exhibit deficiencies in interactivity (e.g., lack of sufficient interactivity options for users), handling attributes/class imbalance (e.g., no samples of large birthmarks), producing believable and high-quality multi-manipulations, as well as in translating audio/image/video to video/speech manipulations. Moreover, creating a taxonomy of audio/facial attributes in A/FAE and A/FAM will greatly help the people in the field.

4.3.3. Real-Time Deepfakes Generation

Advancements in deepfake technology have empowered both malicious attackers and non-malicious users to manipulate faces during live streaming, broadcasting, gaming, entertainment, or videoconferencing. Such real-time facial audiovisual content modification is achieved via the use of AI-powered deepfake filters [481], e.g., increasing the prevalence of smiles or changing eye gaze. Nevertheless, generating high-fidelity and high-precision face swaps or reenactments remains a persistent challenge in the context of real-time Internet live sessions. More compressed, sufficient, and tailored algorithms should be developed for live deepfake generators in in-the-wild scenarios. Moreover, innovate AI-driven deepfake filters capable of producing high-quality audio and video deepfakes; specifically designed for multi-person or multi-face scenarios; accommodating substantial head movements; and addressing aspects such as higher tone, register, vibrato, diction, and breath support should be designed. Future research in real-time deepfakes generation may also focus on improved GANs and diffusion models, enhancing the temporal coherence in video sequences, few-shot and zero-shot learning with minimal data, and dynamic and adaptive neural networks that can adjust their processing power and complexity in real-time depending on the input data. As discussed in Section 4.12.2, the issue of face tampering in live streaming can be mitigated through a challenge–response procedure.

4.4. Advancing Scalability and Usability in Deepfake Techniques

This section delves into scalability and usability considerations within deepfake techniques. Section 4.4.1 discusses the topic in general, while Section 4.4.2 explicitly talks about audio deepfake detection, which is underexplored in this theme compared to visual deepfakes.

4.4.1. Scalability and Usability Considerations

Scalability and usability are important factors to consider when selecting a deepfake technique. The deepfake and audio/face manipulation generation and detection tools' scalability and usability depend on factors like input audio/face (and/or output) resolutions, sizes, dataset, user interface, technical expertise required, speed, and documentation; thus, researchers and practitioners need to be attentive when choosing the most suitable deepfake method for their study, needs, and available resources. For instance, among generation techniques, CycleGAN [200] is comparatively better than StyleGAN [78] due to ability of handling unpaired data and user-friendliness. A comprehensive study to assist users in selecting a suitable technique based on scalability and usability is currently unavailable.

4.4.2. Scalability-Focused Approaches to Audio Deepfake Detection

The trade-off between efficiency and scalability is a notable challenge for current audio deepfake detectors. Thus, innovative schemes need to be formulated that achieve both the highest accuracy and significant scalability simultaneously. Towards this goal, emphasis should be on creating fast and efficient preprocessing, audio DL models, data transformations, self-supervised learning, dynamic optimal transport, gradient compression, distributed architectures, hybrid processing, accelerated system-on-chip, and parallel algorithms that are capable of handling both labeled and unlabeled data.

Evaluating the impact of scalability and usability considerations in deepfake technology involves examining both the technical and societal implications. Some of the key aspects of scalability considerations are computational power, data availability, algorithm

efficiency, optimization techniques, cloud services, resource costs, energy consumption, real-time processing, accuracy and speed trade-offs, deployment (i.e., integration with platforms), user accessibility, and adaptability to evolving threats. The usability considerations are user interface, ease of use, customization options, and interoperability.

4.5. Deepfake Provenance and Blockchain Solutions

The subsection addresses critical strategies for enhancing the traceability and integrity of deepfake content, focusing on source identification (Section 4.5.1) and the application of blockchain technology (Section 4.5.2). With the burgeoning momentum of blockchain technology and its increasing utilization across various sectors, we find it deserving of a dedicated subsection.

4.5.1. Deepfake Source Identification

Another way to alleviate deepfakes is deepfake content provenance (i.e., deepfake source identification). The content provenance can help common people know the authenticity of the information. An initial attempt was made by Adobe, The New York Times, and Twitter (currently X) via the Content Authenticity Initiative (CAI) [482] to determine details like time, location of the media, type of device used, etc. Also, the Coalition for Content Provenance and Authority (C2PA) [483] has been formed, which will direct the enactment of content provenance for editors, creators, media platforms, publishers, and consumers. More attention should be devoted to the aspect of deepfake provenance. Researchers can develop frameworks using blockchain technologies, distributed ledger technologies, and multi-level graph topology, usable on large social networks for deepfake content traceability. Furthermore, deepfake source identification algorithms, which are multimodal and can take text, image, video, and audio as input modalities in both unconstrained and constrained settings, should be devised.

Addressing the challenge of deepfake provenance requires collaboration and coordination among various stakeholders. The responsibility for implementing and supervising Deepfake Provenance (Source Identification) typically falls on a combination of entities, including research institutions, tech companies, government agencies, law enforcement, independent non-profit organizations, internet social media platforms, as well as users and communities.

4.5.2. Blockchain in Deepfake Technology

Conventional deepfake methodologies fall short in examining the history and origins of digital media or algorithms. Utilizing blockchain technology offers a fresh perspective in the ongoing battle against deepfake manipulation. Blockchain is a decentralized and distributed digital ledger technology that facilitates secure, transparent, and tamper-resistant record-keeping of proceedings and the monitoring of resources securely and independently [484–487]. Few studies have explored blockchain in deepfake technology—this study area is still in its infancy. Future advancements in blockchain, smart contracts, hyperledger fabric, hashing techniques, security methods, and integrity measures in the realm of deepfake technology is crucial for deepfake provenance, modifications made to it, and the authenticity of media assets. These innovations will help in uncovering the user motivations behind deepfake postings/publications and addressing the potential misuse of deepfake generation and detection algorithms.

4.6. Fairness and Credibility of Deepfakes

Recent studies observed that deepfake (detection) models and datasets are negatively biased and yield partisan performance for different racial, gender, and age groups. Certain groups are excluded from correct detection and unfairly targeted, leading to fairness, security, privacy, generalizability, unintentional filtering of harmless contents, and public trust or opinion issues. Furthermore, a study conducted in [488] revealed that racial adjustments in videoconferencing sessions resulted in higher speaker credibility when

participants were perceived as Caucasian. It is imperative to assess and discern bias (i.e., favoritism or unfairness absence) in deepfake technology for massive roll-out and commercial adoption. Little research has been conducted on ensuring fairness in deepfake technology, and there is a need to fill this gap. Towards this goal, we can explore new loss functions that are agnostic to demographic factors, non-decomposable loss with large-scale demographic and non-demographic attributes, and reducing pairwise feature correlations. In addition, advanced training algorithms can be developed to specifically isolate or remove demographic features, e.g., deep generative models to eliminate gender information from input samples before feeding them into feature extraction for classification. Likewise, fairness penalties and risk measures can be incorporated into the learning objectives.

Specific features also contribute to the lower fairness of deepfake detectors; hence, multi-view features-based systems comprising multiple micro/local and macro/global mechanisms should be constructed. Also, collecting and publicly releasing more balanced, unbiased, and diverse demographic databases will recuperate the fairness issue. Undoubtedly, substantial strides are necessary for the continued advancement in creating balanced datasets and fostering fairness within deepfake frameworks. Besides, the social media and video-conferencing companies should introduce a feature for transparency, e.g., full disclosure of face ethnicity alteration (i.e., digital modifications of the perceived ethnic characteristics of a person's face) to all parties on the post or video call (i.e., displaying a red mark).

All in all, evaluating the fairness and credibility of deepfakes necessitates a multi-faceted approach combining quantitative and qualitative methodologies. Quantitative approaches such as statistical analysis (e.g., detection accuracy, false negative rate, confidence scores, and types of deepfake) and data analysis (e.g., deepfake distribution across different platforms and contexts, geographical analysis, and temporal evolution) can be applied. Qualitatively methods such as such as user perception studies (e.g., surveys and interviews), content analysis (e.g., categorizing themes, purposes, context, and potential impacts), and expert evaluation (e.g., involving experts from ethics, law, journalism, and technology) could be adopted to understand the societal impact and ethical considerations. The integrated approaches that merge quantitative and qualitative methods ensure comprehensive analyses that addresses both technical reliability and the broader societal implications of deepfake technologies.

4.7. Mobile Deepfake Generation and Detection

Mobile devices have become our natural extension, which is expected to expand in coming years with high interconnectivity. Deep neural network-based deepfake generation and detection, while highly accurate, is often impractical for mobile platforms or applications due to complex features, extensive parameters and computational costs. For instance, models like temporal convolutional networks [489], denoising diffusion probabilistic models [289], adversarial autoencoders [490], and Stack GANs [491–493] used in deepfake generation and detection can demand substantial computational power and memory, which are often beyond the capabilities of mobile devices. Light-weight [494] and compressed deep learning-based [495] detection systems, which are yet efficient and deployable on mobile and wearable devices, will be instrumental in mitigating the impact of misinformation, fake news, and deepfakes. There are huge opportunities to devise frameworks for both generating and detecting deepfakes on mobile devices, particularly designed for online meetings and video calls.

4.8. Hardware-Based Deepfake Detectors

Everyday, millions of hours' worth videos are exchanged or uploaded online; therefore, to thwart fake news, mis/disinformation, and deepfakes, researchers and professionals can formulate hardware technologies that can detect such contents proactively. For example, developing dedicated media processing chips/systems (e.g., ASIC (application-specific integrated circuit) [496] and TPMs (Trusted Platform Modules) [497]), which directly scru-

tinize content for their authenticity. Hardware accelerators exhibiting parallelism, soft cores emulation, energy efficiency, rapid inference, and fileless processing capabilities are poised to drive real-time and better-quality deepfake detection in unconstrained situations. Any such innovation is poised to revolutionize the field, spawning a plethora of new applications.

4.9. Part-Centric Strategies for Deepfake Detection

This subsection delves into part-centric approaches, with a specific focus on Sections 4.9.1 and 4.9.2, which specifically address the detection of part-based face and part-based audio deepfakes, respectively.

4.9.1. Part-Based Deepfake or Face Manipulation Detection

Most of the prior deepfake and face manipulation detection systems employ whole face image/frame. Nonetheless, often certain face parts are cluttered or redundant, leading typically to inferior performances. To overcome such issues, face part-based techniques can be designed, for example, selecting particular facial components (e.g., right eye or mouth) for deepfake detection [498]. The part-based methods to locate facial attributes and extract features can be performed using auxiliary part localization (e.g., existing facial part detectors) and end-to-end localization (e.g., unified framework to locate facial part and classification) algorithms [453].

4.9.2. Partial Audio Deepfakes

Publicly available data samples within the domain of deepfake technology are predominantly concentrated on scenarios where the spoken words are entirely either real or deepfake. The real-word scenarios where only a small part like a short section or a single word is manipulated (i.e., partial deepfake [40]) are hard to detect. It is worth noting that partial deepfakes might not trigger false alarms in automatic speaker verification systems. However, they can still easily alter the meaning of a phrase. The continuous estimation of labels at the segment level (i.e., which portions are real and which are fake) study will help to comprehend the rationale behind a classifier's specific decision. The demand for extensive partial audio deepfake datasets in various languages emphasizes the necessity for segment-level labels and metadata for better analysis and detection efforts.

4.10. Multi-Person Deepfake Databases and Algorithms

The majority of publicly available deepfake datasets are primarily composed of scenarios where a single person/face/audio is manipulated or deepfaked within a given sample, or the sample itself contains only one face/person/audio. Similarly, a systematic analysis shows that the majority of existing methods are concentrated on single-person deepfake scenarios and fail under multi-person or multi-face deepfaked situations. However, multi-person deepfakes (i.e., where deepfakes of multiple individuals coexist in a sample) mirror real-world scenarios accurately. Therefore, wide-reaching deepfake databases featuring multiple people (i.e., multi-faces and/or multi-speaker) in uncontained settings with known and unknown numbers of manipulations should be created. To this aim, the Openforensics [126] and DF-Platter [161] datasets were recently created. No publicly available audio deepfake datasets currently feature multiple deepfaked speakers in the same sample with varying real-world conditions, to the best of our knowledge. Furthermore, formulating sophisticated algorithms for creating and detecting multi-person audio and video deepfakes will significantly propel the evolution of deepfake technology.

4.11. Occlusion-, Pose-, and Obscured-Aware Deepfake Technology

Current deepfake generators and detectors struggle to effectively manage deepfakes with occlusion, pose, obscurity, and mask variations. Namely, deepfake detectors may have limitations in identifying manipulated audios/faces when they are obscured by facial masks, heavy makeup, vocal peripherals, vocal cavity, phonology, background acoustics,

and interference, or when only a partial portion of the speech/face is manipulated. Future research in this area could focus on directions such as occlusion and/or pose detection and inpainting, contextual understanding, 3D modeling and reconstruction, differentiable rendering, multi-view learning, pose normalization, partial convolution, feature disentanglement, adaptive feature extraction, and hierarchical feature representations, which can focus on available information while compensating for missing parts.

Similarly, deepfake generators often lack consistency and quality in their outputs. In addressing such challenges, recently, a few frameworks (e.g., [223,499]) have been proposed. Further emphasis should be placed on the advancement of deepfake generators that possess occlusion-, pose-, and obscurity-awareness; high resolution, consistency in identity-related and -unrelated details; real-time functionality; and high fidelity. Additionally, generators should be lightweight and adept at processing/generating diverse qualities for arbitrary source and target pairs. Also, the availability of public databases containing partial, occluded, and obscured deepfakes, along with manipulated speech/face samples, is currently limited.

4.12. Instantaneous and Interactive Detection Methods

This subsection explores the field of live deepfake detection, with Section 4.12.1 discussing instantaneous detection in general and Section 4.12.2 focusing on participatory deepfake detection.

4.12.1. Real-Time Deepfake Detection

Certain challenges persist within the existing deepfake detectors, which require careful attention and resolution, e.g., many algorithms lack real-time applicability. Namely, highly precise detection algorithms require longer inference time. Social media and online services (e.g., voice or face-based authentication in online banking) will continue to expand, as does the threat of deepfakes. The irreversible impact may materialize before the detection or realization that the content is manipulated or fake. Consequently, the call is for real-time, highly accurate deepfake detectors in real-world applications. Special emphasis should be given to future techniques that operate in real-time without being resource-intensive. Novel real-time deepfake detector schemes should exhibit reliability, robustness, swift processing, and effectiveness across various platforms and situations. With this objective in mind, researchers may delve into compressed DNN (deep neural network) modeling, lightweight dynamic ensemble learning, and contextual learning.

4.12.2. Challenge–Response-Based Deepfake Detection

Challenge–response deepfake detection can be employed in online audio calls, video calls, or meetings, where users are required to perform specific tasks/actions to be labeled as a genuine content/user for verification via the proper execution of random instructions in order. This could also prevent bots from performing automated deepfakes in live online calls/meetings. This method may be user-unfriendly, with high computational costs and low acceptability. Innovative, effortless response, and user-friendly techniques need to be devised, e.g., the system can pose challenges triggered by factors such as lack of change in voice tone [500], movement, or facial expressions. The deepfake attacker can easily sidestep a single challenge; therefore, a system can implement a series of challenges, as circumventing an entire series of challenges is notably much harder.

4.13. Human-in-the-Loop in Deepfake

Deepfake and audio/facial attribution manipulations are continuously attaining a level of sophistication that surpasses ML-based detectors' capacity [501,502]. The battle between audio/visual forgery and multimedia forensic continues without resolution. Thus, rather than automated AI/ML deepfake systems without human involvement, including humans in both the training and testing stages of devising deepfake generation/detection algorithms would produce better results because of a continuous feedback loop. We can

term it as the human-in-the-loop deepfake approach, which can leverage the effectiveness of intelligent ML automation while remaining responsive to human feedback, knowledge, and experiences via Human–Computer Interaction (HCI). To the best of our knowledge, there is currently no human-in-the-loop deepfake detection method. The question of how to embody meaningful and useful human knowledge, attributes, and interaction into the system still remains open. Feedback from someone who cannot recognize deepfakes will be valuable in a human-in-the-loop approach to improving deepfake systems. Such feedback may help in various ways, such as highlighting particular weaknesses in current algorithms, and insights from non-experts guide developers to improve accuracy and robustness, enhancing the diversity of training data, improving user interface design, and addressing real-world detection challenges and biases across different user groups and contexts.

4.14. Common Sense Audio/Face Manipulations and Deepfakes

The predominant phoniness resides in the visual texture and timbre of audio/sound. Humans can very easily identify the common-sense audio/face manipulations and deepfakes, e.g., human face with horns, one eye on the chin area, horizontal nose on the forehead, ears at the position of mouth, malformed face, artificial intonation, etc. However, deepfake detection techniques cannot identify them as they are devoid of basic common sense. There is a scarcity of public databases with common sense audio, visual deepfakes, and face manipulations.

4.15. Large-Scale AI-Generated Datasets

Examining published literature reveals that research on detecting AI-synthesized deepfakes often utilizes custom databases derived from various generative [503,504] and diffusion [505,506] deep models. Since there is no consensus on the choice of parameters and datasets, the bulk of published studies exhibit diverse performances under GAN and diffusion samples, as the quality of AI-generated samples varies and remains largely unknown. To foster preeminent advancements, the community ought to create several all-inclusive public AI-generated audio and video deepfake datasets, housing samples of different manipulations, qualities, ages, ethnicities, languages, tones, educations, and backgrounds with both constrained and unconstrained scenarios.

4.16. Deepfakes in the Metaverse

The metaverse [507] will soon be integral to our lives, reshaping how we learn, live, work, and entertain while enabling imaginative experiences beyond the constraints of time and space. Harnessing deepfakes in positive manners in the metaverse can result in extended reality applications that are both hyper-realistic and profoundly immersive. But, in the metaverse, malicious actors might employ deepfakes to deceive and manipulate others, potentially resulting in diverse fraudulent activities [508]. Current deepfake detection methods, originally designed for 2D deepfakes, are inadequate in the metaverse as metaverse deepfakes are primarily based on 3D rendering and modeling. To counter XR, AR, VR, and MR (i.e., the metaverse) deepfakes effectively, it is required to develop intricate and multi-layered systems of robust and preventive safeguards. Future techniques for generating and detecting metaverse deepfakes demand lower complexity, real-time capability, and high resolution.

4.17. Training and Testing Policies

Collaborative efforts between research and professional communities are essential to formulate comprehensive training and testing policies for deepfake technology. It is advisable to incorporate both a fixed policy and a more flexible one. The strict or fixed policy can be utilized for meaningful comparisons of different deepfake frameworks as well as a basis for benchmarking evaluations. For instance, the fixed policy could include standardized datasets (e.g., DFDC); uniform and well-adopted evaluation metrics (e.g., accuracy and area

under the receiver operating characteristic curve); and benchmarking protocols including the environment setup, hardware specifications, and procedures (e.g., GPU model and software environment). The flexible or adaptable policy can be swiftly useful for future simple, complex, compressed, larger, or bigger frameworks. The flexible policy may be designed to handle, for example, varied frameworks (e.g., personal choice-based DL methods among GANs, RNNs, and CNNs), new or emerging datasets, and innovative metrics (e.g., evaluation metrics that can evaluate both robustness and detection/generation). The training and testing policies, whether fixed or flexible, must explicitly outline procedures and considerations for generating and detecting individual audio, video, and text deepfakes, as well as joint audio–visual and text-to-audio/multimedia deepfakes.

4.18. Ethical and Standardization Dimensions in Deepfake Technology

This subsection explores two critical dimensions of deepfake technology, i.e., the efforts to standardize its generation and detection (Section 4.18.1) and the ethical implications thereof (Section 4.18.2).

4.18.1. Standardization of Deepfake (and Audio/Face Manipulation) Generation and Detection

Deepfake standards serve as the overarching principles for generating, collecting, storing, sharing, and evaluating deepfake or digitally manipulated audio/face data. There are no international standards yet for deepfakes, although some attempts are underway by the International Organization for Standardization (ISO) [509], European Telecommunications Standards Institute (ETSI) [510], and Coalition for Content Provenance and Authenticity (C2PA) [483], to name a few. It is crucial to set accurate standards to fully maximize the potential of deepfake technologies. As high-level decision-makers and government agencies assess the trade-offs between the deepfake risks and the convenience of digital rights, they must recognize the necessity of implementing proper deepfake standards for privacy, vulnerability, development, sale, and certification.

4.18.2. Ethical Implications of Deepfakes

The ethical concerns surrounding deepfakes are continuously growing [511], as they can be utilized for blackmail, sabotage, intimidation, incitement to violence, ideological manipulation, and challenges to trust and accountability. To foresee the ethical implications, it is vital to grasp its current state, limitations, and potential opportunities. While technological aspects have garnered significant attention, the ethical dimensions of deepfake technology have received comparatively less attention. Proactively developing and adopting ethical standards will help mitigate potential risks. To achieve this objective, a thorough study on the ethical implications of deepfakes across diverse applications and domains should be undertaken for informed ethical developments, for instance, investigating questions like 'what kinds and levels of harm can rise?' and 'what steps can be taken to minimize these harms?'. Ethicists, information technologists, lawmakers, communication scholars, political scientists, digital consumers, business networks, and publishers will have to join forces to develop "Ethical Deepfake" guidelines akin to "Ethical AI" guidelines. For example, ethical deepfakes in business and marketing should be transparent (i.e., content source), non-deceptive (i.e., content is not real), fair ((i.e., respecting the rights of third parties), and accountable (i.e., providing the option to consumers to opt out of deepfake content if desired).

4.19. Regulatory Landscape for Deepfake Technologies

Here, Section 4.19.1 concentrates on the complexities of protecting personal privacy amidst the proliferation of deepfakes, as well as examining the various legal liabilities arising from their misuse, while Section 4.19.2 studies legislative efforts worldwide aimed at mitigating risks and ensuring responsible use of this technology.

4.19.1. Deepfakes Privacy and Liability

Privacy: The surge of deepfakes raises grave concerns regarding the compromise of personal privacy, data privacy, and identity security [512]. Consequentially, different techniques to enhance individual's privacy and security in audio/facial samples or deepfakes have been suggested such as pixelation [513], blurring [514], masking [515], background noise or sprechstimme addition [473], cartoons [516], and avatars [517].

On the contrary, deepfakes themselves have lately been regarded as a means to bolster privacy and security, e.g., modifying, swapping, and manipulating the original face style with a machine-generated face/avatar and attributes, respectively. Additionally, to safeguard the privacy of audio/face samples used in training of 'entire audio/face synthesis' models, approaches like watermarking [518], obstructing the training process [519], or removing identity-related features first before use [520] have been introduced. Despite their utility, the mentioned approaches frequently show suboptimal generalization to unfamiliar models and introduce unwanted noise to the original sample. Wide-ranging studies are necessary to evaluate the efficacy of deepfakes in concealing individuals' identities and to quantify the extent to which privacy-enhancing deepfakes impact the performance of speech/face recognition systems.

Text-guided audio/face sample editing strategies [521] should be developed to empower users to carry out deepfake privacy procedures according to their intentions and level of desired privacy or security. Similarly, social networks and deepfake apps should provide users with the option to appear only in manipulated audio/visual samples/posts they have explicitly approved as well as to choose whether their face/audio needs to be deepfaked.

Liability: Deepfakes pose several significant liabilities such as intellectual property infringement (e.g., using deepfake technology owned by others without proper authorization), invasion of privacy (e.g., unethical exploitation of people's images to create deepfakes or the illicit sale of such manipulated content), defamation liability (e.g., utilizing deepfakes to disseminate false information, potentially tarnishing someone's reputation), data breach liability (e.g., a compromise of data protection and privacy through unauthorized disclosure, substitution, modification, or use of sensitive identity data for deepfakes), and unfair competition liability (e.g., promotional marketing materials featuring a deepfaked spokesperson). Addressing these concerns demands a comprehensive framework that navigates the intricate legal and ethical landscapes surrounding the deepfake technology.

4.19.2. Deepfake Legislation

The current regulatory system is inadequate in addressing the potential harms associated with deepfakes. Legislators are contemplating new legislations to grapple with deepfake technologies, which could significantly impact individuals, nation, companies, and society as a whole. For example, the states of Virginia and New York have respectively criminalized nonconsensual deepfake pornography and non-disclosure of deepfake audio, visual, or moving-picture content [522]. Policymakers across the globe should put more dedicated efforts to regulate and create criminal statutes, as numerous regulatory proposals are still in the initial phase. The new and effective legislations and regulations should foresee, accommodate, and mitigate current and future deepfake harms without impeding technological development or stifling expression. It will require nuanced collaboration between civil society organizations, technology companies, and governments to formulate adaptable regulatory schemes capable of addressing the evolving nature of deepfake threats.

4.20. True-to-Life Audio Deepfakes

The field of audio deepfake generation and detection is gaining momentum [461,523]; however, the existing body of literature on visual/image deepfakes far surpasses that of audio deepfakes. Audio deepfake detection is still problematic, particularly in discerning synthetic voice tracks created through advanced techniques and samples derived from open-

set scenarios. Moreover, many detectors use synthetic speech recognition features, which compromise accuracy and leave them susceptible to adversarial examples and realistic counterfeit audio [524]. More advanced audio deepfake generators should be formulated that can produce more authentic audio tracks, incorporating diverse background noises and contextual variations. Research communities should create extensive public audio deepfake databases that cover diverse languages, background noises, codings, and a variety of vocal characteristics including tenor, bass, mezzo-soprano, baritone, contralto, countertenor, falsetto, whisper, narrative voice, families of synthetic speeches, and different age groups. These databases should also include recordings made using PA systems and mobile devices to ensure a comprehensive representation for effective research in audio deepfake technologies. Additionally, large-scale deepfake samples, subject to simultaneous language conversion and manipulation, should be released publicly with accurate labels.

4.21. Effects of Non-Speech Intervals in Deepfakes

Non-speech intervals naturally occur in spoken language and can also assist in distinguishing between authentic and manipulated speech. Attackers can adroitly eliminate non-speech segments to circumvent audio deepfake detectors. Exhaustive investigations are required to gain insights of the generation or alteration of non-speech segments, as not all non-speech frames/types (e.g., breathing sounds) in different languages have been explored. In a similar vein, it is crucial to conduct research on the impacts of varying durations of non-speech intervals in deepfakes across different languages, investigating their correlations within languages and across various languages and modalities.

4.22. Future Deepfake Datasets

This section discusses cutting-edge deepfake datasets across three key dimensions: diverse main categories of audio deepfakes (Section 4.22.1); varied audio datasets with different attributes, e.g., styles, single and multi-person (Section 4.22.2); and a comprehensive collection of audiovisual deepfake media spanning a wide range of subjects, scenarios, languages, and contexts (Section 4.22.3), akin to the breadth of content found in an encyclopedia. Each part explores advancements in technology and methods that will catalyze the advancement of the deepfake field.

4.22.1. Diverse Audio Deepfakes

Publicly available audio deepfakes samples or comparative datasets were mainly created using the TTS (text to speech), SS (speech synthesis), and VC (voice conversion) methods [47]. Studies on recent TTS, SS, and VC methods, which need minimum training data to generate cutting-edge deepfake samples, should be conducted. A fresh effort should be initiated to collect and generate manipulated audio tracks using software-based real-time voice changers, voice cloners, vocal synthesizers, language deepfake translations, digital signal processing-based techniques, next-generation neural networks, and hardware-based voice changers. Furthermore, examining deepfake detectors and speech recognizers' vulnerabilities against the abovementioned manipulation techniques with different types of scenarios (e.g., background noises, vocal cavity, vocal peripherals, fixed microphone, or access control) will aid next-generation developments.

4.22.2. Heterogeneous Audio Datasets

Current, deepfake audio datasets predominantly feature samples from one to three languages and include only one or two types of deepfake instances [12]. As a result, detectors struggle to generalize and tend to overfit [332,473]. Developing heterogeneous genuine audio and deepfake datasets with multiple languages, manipulation techniques, speaking styles, scenes, partial and full manipulated tracks, same single track containing multiple types of audio deepfakes, and relevant factors will advance audio deepfake technologies and applications, including the robustness and versatility of models.

4.22.3. State-of-the-Art Encyclopedic Deepfake Datasets

While numerous deepfake datasets (e.g., [121,330]) are publicly accessible for research and development, they come with several limitations. For instance, many datasets lack coverage of the latest deepfake audios and/or videos crafted through state-of-the-art (SOTA) techniques available on various platforms. The datasets lack diversity in quality, age, and ethnicities, with deepfakes showing noticeable and/or visible disparities from those circulated online and imbalanced class distributions. Hence, deepfake detectors tested on or developed using these datasets encounter challenges related to overfitting and generalization issues. Forthcoming encyclopedic databases for audio and video deepfakes should be meticulously crafted to include SOTA-generated deepfakes; multi-language, different quality, demographic data; realistic context and scene representations; various audio/facial attribute manipulations; samples collected in the wild; full and partial manipulations; multimodal forgeries; distinct resolutions; superior-quality synthetic and morphed samples; and adversarially perturbed, postprocessed (i.e., postprocessing operations on deepfakes remove artifacts to look/sound more lifelike), and metadata. Such extensive and diverse datasets contribute to the development of sophisticated countermeasures against the evolving landscape of real-world deepfakes. Additionally, although 'not safe for work' deepfake audios and/or videos are accessible online, public datasets containing pornographic deepfakes in the wild are not widely available.

4.23. Imitation-Based Audio Deepfake Detection Solutions

The prevailing detectors largely address synthetic-based deepfakes, which leaves imitation-based audio deepfake detection understudied. Identifying audio deepfake content generated via the imitation-based technique is challenging owing to the strong similarity between original and fake audio samples [525]. This conspicuous gap and issue underscore the urgency of redirecting analytical efforts towards the refinement of imitation-based deepfake detectors, including identifying subtle differences in fake speech imitation features as well as algorithms the can detect both synthetic- and imitation-based deepfakes.

4.24. Proactive Deepfake Defenses

In deepfake passive defenses, detectors undergo retrospective training to discern manipulated samples. This methodology functions as post-event forensics but proves inadequate in preemptively thwarting the dissemination of disinformation. Therefore, initiatives for deepfake proactive defenses have been put forward. This methodology operates as a preemptive forensic analysis, aiming to obstruct the generation of deepfakes and promptly identify them before they swiftly spread across diverse platforms. Proactive approaches bear some resemblance to anti-counterfeit measures. Proactive methods enhance forensic tools in order to fortify overall deepfake defense. Some of the proactive deepfake countermeasures are watermarking [518] (i.e., tracing the copyright of audiovisual samples by examining embedded visible or hidden watermarks or embedding anti-deepfake labels in speech/facial features to detect tampering by indicating the presence or absence of the labels), adversarial examples [526] (i.e., injecting adversarial perturbations into real data creates distorted deepfake outputs, which may be easily noticeable to human observers), and radioactive training data [527] (i.e., infusing data with subtle alterations ensures that any model trained on it will produce deepfakes with a unique and identifiable 'fingerprint'). Proactive defense techniques, according to recent studies, demonstrate specific limitations. For instance, they can be comfortably sidestepped by basic audio/image transformations to the original perturbated samples, reconstruction of the original sample from deformed deepfake outcome, and eradicating the watermarks or fingerprints from the samples. Furthermore, many current proactive methods incur high computational costs. Subsequent endeavors should focus on developing holistic deepfake proactive solutions by incorporating techniques such as watermarking, adversarial examples, and radioactive training in a unified approach. Future proactive defenses should be designed to generalize across unseen deepfakes and withstand tampering and postprocessing, which can be achieved

using ensemble watermarking (e.g., integration of invisible and visible watermarks or robust and fragile watermarks) and blockchain methodologies. There is a shortage of deepfake proactive defense mechanisms for audio and multimodal samples.

4.25. Explainable Deepfake Detectors

The reliability of a deepfake detector stands on the pillars of transferability, robustness, and interpretability. The interpretability (also referred to as explainability [528–532]) involves the capability to understand, interpret, and offer transparent insights into the processes and decisions made by a deepfake detector, thereby fostering trust among users, researchers, and policymakers in comprehending the intricacies of the detection process. A plethora of solutions have emerged to confront challenges related to transferability and robustness, but the aspect of interpretability has been relatively underexplored. Due to the lack of explainability in deepfake detectors, their outcomes struggle to instill trust among the public and face challenges in gaining acceptance in critical scenarios, such as legal proceedings in a court. Deepfake detectors based on deep learning lack the ability to offer human-understandable justifications for their outputs, primarily because of their black-box nature. Present speech/face manipulation or deepfake detection frameworks solely offer a categorization label (e.g., pristine, deepfake, or digitally manipulated), a probability score indicating the likelihood of fakeness, or a confidence percentage for audio or video content being fake, without accompanying explanations and reasons for the detection results. Additionally, deepfake detectors are unable to discern whether the speech/face manipulation or deepfake was carried out with benign or malicious motives, and which tool was utilized to generate it.

A few explainable deepfake detection schemes have been proposed, often employing attention mechanisms like heatmaps to highlight manipulated regions in audio or video. The dependability and transparency issues in deepfake detection are yet to be fully resolved, requiring answers to questions such as the following: why is the audio or video classified as deepfake?; on which portion and by what criteria does the detection algorithm identify the audio or video as fake?; how can logical reasoning be efficiently integrated to improve the interpretability of multimodal deepfakes?; how and which aspects of psychological knowledge can enhance the generalization capability of transparent deepfake detectors?; and in what ways can future explainable methods be crafted to democratize deepfake detection for individuals not well-versed in AI?

Future efforts should prioritize transparent end-to-end frameworks that incorporate multiple cues (e.g., physiological and physical indicators), since a singular cue alone cannot encompass the diverse range of artifacts and attributes. Additionally, emphasis should be placed on exploring multimodal explainable approaches and developing universally acceptable explainability indicators or measures. To cultivate proficient interpretable deepfake detection systems, leveraging dynamic ensembles of techniques—such as knowledge distillation, multimodal learning, layer-wise relevance propagation, neural additive models, and fuzzy inference—can be advantageous. Deepfake interpretability can also be achieved through interdisciplinary research, involving a team of experts in social networks capable of analyzing posts, re-posts, and comments, along with fact-checkers, data mining specialists, and computer vision and AI scientists.

4.26. Multitask Learning-Based Deepfake Technology

The paradigm of multitask learning in machine learning involves training a framework to tackle several tasks simultaneously. Multitask learning has been empirically shown to yield superior performance when compared to single-task learning [533]. The future research emphasis should be on developing unified deepfake generation and detection frameworks rooted in multitask learning. These frameworks should jointly handle various factors, including different resolutions, quality variations, occlusions, demographic classes, poses, and expressions, to drive progress in the field. Towards this objective, schemes can meticulously design loss functions, including but not limited to face prior loss, perceptual

loss, adversarial loss, style loss, smooth loss, pixel loss, temporal loss, and inter-modality loss. Extracting features from the forgery location consequentially contributes to the success of deepfake detection [534]. Therefore, creating deepfake detection methods capable of simultaneously performing deepfake detection and forgery location could enhance generalization capability and bolster resilience against adversarial attacks and postprocessing tricks, such as resizing or pitch shifting.

4.27. Computational Efficiency

The traditional deepfake approaches based on computer graphic or digital signal processing are relatively costly and time consuming, with low processing speed and high computational complexity. Comparatively, deep learning (DL)-based deepfake generation and detection frameworks are easier to use and more computationally efficient. For instance, a pre-trained GAN model for deepfake generation can perform quickly and efficiently during inference compared to traditional methods involving facial feature modeling, texture mapping, and rendering pipelines. However, despite their advantages, these DL-based frameworks still cannot be used widely in practical scenarios as they mainly focus on accuracy improvements. In particular, for good quality of deepfake generation, DL methods require larger databases. To reduce the computational complexity during training, techniques such as data distillation, one-shot learning [535], and few-shot learning [536] should be fully explored. In both detection and generation, by considering image/video/audio resolution, dataset sizes, deepfake types, and model complexity, efforts should be directed towards devising proficient light-weight schemes with cycle-consistency loss and/or federated learning and edge-cloud systems.

4.28. Battle of Deepfake Detection: Humans versus Machines

Deepfakes and manipulated multimedia content, if widely circulated, can have a detrimental influence on society at large. Human deepfake detection capabilities exhibit significant variation, influenced by diverse factors such as context, audio or visual modalities, quality, realism, coherence, and other variables. Few studies (e.g., [501,537,538]) have delved into the comparative analysis of human proficiency against machine in identifying deepfakes. There is no consensus in findings, but existing studies indicate the following: (i) ordinary people struggle and need more attention to spot deepfakes than experts; (ii) people are better at spotting fake speeches when they have both audio and visual elements; (iii) automated ML deepfake detectors are slightly more effective with blurry, noisy, grainy, or very dark deepfakes (but it is well known that automated detectors show limitations when they encounter novel deepfakes not used in their training); (iv) human detectors and ML deepfake detectors make different types of errors; (v) synchronous collective performance (of groups of individuals) can match or surpass individual subject and ML deepfake detector detectors accuracy; (vi) humans can easily detect deepfakes in political videos; (vii) financial incentives have limited impacts on humans to improve the accuracy; (viii) in ML deepfake detectors, combining both sight and sound leads to higher accuracy than that of humans; (ix) computers and people perform similarly in detecting audio deepfakes; (x) native speakers are usually better at catching fakes than non-native speakers; (xi) individuals are more vulnerable to video deepfakes compared to audio ones; (xii) most studies are on English and Mandarin language deepfakes; (xiii) language does not affect detection accuracy much; (xiv) shorter deepfake clips are easier to identify; (xv) increasing awareness of deepfake existence marginally enhances detection rates; (xvi) difficulty of the detection by human subjects rises when background/context knowledge is eliminated; (xvii) detection accuracy relies on human knowledge and expertise; (xviii) impairing visual face processing impairs human participants but not ML-based detectors; (xix) human individual and group biases (e.g., homophily bias and heterophily bias) play a vital role in detection capabilities; (xx) overly confident subjects in their own detection abilities exhibit reduced accuracy in deepfake detection; (xxi) demographic traits strongly affect deepfake detection (e.g., male, non-Caucasian, and young individuals are more accurate); (xxii) audio

deepfakes crafted using text-to-speech methods are difficult to distinguish than the same deepfakes with voice actor audio; (xxiii) human subjects place more emphasis on 'how something is said' rather than 'what is said'; (xxiv) examining consistency and coherence in texts is more advantageous to detect text deepfakes.

In essence, machines currently outshine humans in detecting deepfakes due to their ability to discern subtle details often overlooked by human observers. Human and machine perceptions differ, but both can be fooled by deepfakes in unique ways. Collaboration between academics and professionals in digital media forensics, social psychology, and perceptual psychology is crucial for a comprehensive understanding. The focus of future endeavors should be on devising novel schemes to fuse human deepfake ratings (e.g., mean opinion scores), and comprehensive, large-scale assessments of human versus machine performance in the realm of deepfake technology, considering deepfake types, modality (audio, visual, multimodal, and text), quality, styles, resolutions, color depth, codec, aspect ratio, artifacts, number of people in deepfake samples, real-world situations, languages, individual, crowd, and demographic traits. Conducting future studies will aid in identifying the factors that optimize human deepfake detection performance, if any reciprocal learning between humans and machines exist, and effective ways to integrate human cognition and perception into deepfake detection. There is a need for systematic qualitative and quantitative research to unravel the encoding of modality-specific unique characteristics in deepfakes. There has not been any assessment of human versus machine performance specifically for large-scale partial deepfakes. The findings from these studies and works could contribute to designing tailored cybersecurity training programs aimed at enhancing the human detection of deepfakes, fake news, and misinformation.

4.29. Navigating Security Threats: Adversarial Attacks and Audio Noises

Here, we discuss addressing the complexities of navigating security threats posed to deepfake detectors by adversarial attacks (Section 4.29.1) and the intricacies of audio noise (Section 4.29.2).

4.29.1. Vulnerability to Adversarial Attacks and Anti-Forensics

Recent research investigations reveal that deep neural network-based deepfake detection techniques, owing to their inherent flaws, lack robustness against adversarial examples [44] and anti-forensics [539]. Adversarial examples are samples intentionally perturbed (e.g., by adding imperceptible noise) to mislead deepfake detectors [540]. Anti-forensics (aka counter-forensics) is the deliberate act of eliminating or obscuring deepfake evidences in an effort to diminish the efficacy of forensic investigation methods [539]. In anti-forensics, an adversary can modify original deepfakes through different procedures such as resizing, compression, blurring, rotation, light variations, noisy effects, motion blur, and quality degradation. This impedes forensic tools from extracting essential clues about manipulations, falsifications, source devices, etc. Adversarial examples and anti-forensics can be strategically utilized to effectively undermine the classification accuracy of automated deepfake detectors.

There is currently a lack of substantial work focused on robust deepfake detectors against adversarial examples and anti-forensics. Deepfake detection methods experience a drastic decline in accuracy when confronted with novel adversarial attacks and counter-forensics techniques. The trajectory suggests that this trend is poised to grow in the near future. More resilient deepfake detection schemes should be developed, which have the ability to combat not only conventional deepfakes but also those enhanced with adversarial examples and anti-forensic measures. In pursuit of this goal, the strategic design of different filtering, photoplethysmographic features, multi-stream processing, deep reinforcement learning, and incorporating noise or adversarial layers in the detection network holds significant promise. Furthermore, there is a need to design next-generation deepfake detectors that are robust against social media laundering. Social media laundering (i.e., to conserve network bandwidth and safeguard user privacy, videos typically undergo

metadata removal, downsizing, and heavy compression prior to being uploaded on social platforms) eliminates traces of underlying manipulation, leading deepfake detectors to misclassify deepfakes as real.

Also, the red team can pioneer the development of next-generation adversarial attacks and anti-forensics. For example, few deepfake detectors rely on physiological signals, like heart rate extracted from videos or respiration rate extracted from audio, as potent indicators of falsification. This is because GAN techniques fail to preserve the intricate color variations associated with the heart rate/respiration rate signals. To outsmart new detectors, an attacker can design a method for fake videos/audios to contain a normal heart rate/respiration rate signal by introducing sequential color variations in frames/samples. Moreover, current anti-forensics primarily focus on aural presence/appearance artifacts (e.g., distinct frequency spectrum generated by deepfake generators) but overlook their impact on perception (e.g., deep speech/face features). Future advancements may involve the joint removal of both artifacts and perceptual elements for more effective detection. Attackers can design advanced audio, visual, and multimodal frameworks to simultaneously eliminate fake traces by both aural presence/appearance artifacts and perceptual intricacies.

4.29.2. Effects of Noise on Audio Deepfake Detectors

Research has indicated that natural noises (e.g., rain, traffic, machinery, reverberation, wind, thunder, non-stationary signals, babble noise, white noise, and pink noise) and electrical noises (e.g., electrical interference, channel distortions, and cross-talk) can hamper speech recognition system performance [541–544]. Noise is typically defined as 'any arbitrary, unwanted, or irrelevant sound or (electrical) disturbance that interferes with the desired signal or information'. Present audio deepfake detectors struggle with both natural and electrical noises, thereby creating opportunities for attackers to effortlessly deceive them by introducing such noises. Large-scale studies are yet to explore the precise impacts of natural and electrical disturbances on automated audio deepfake detection frameworks. Studies that consider samples recorded in both indoor and outdoor settings with various noise conditions would provide valuable insights for future research, e.g., developing robust deepfake audio detectors applicable in real-world noise contexts.

4.30. Cybersecurity Attacks using Deepfakes

Deepfakes can be utilized as a tool in different types of cybersecurity attacks [545], such as malware/ransomware attacks (i.e., deepfake messages as a vector to deliver malware/ransomware by tricking victims into interacting with malicious content disguised as authentic communications), social engineering (i.e., lifelike deepfakes of trusted individuals exploiting the trust of unsuspecting victims and gaining unauthorized access to sensitive information or systems), phishing attacks (i.e., audio or video realistic deepfakes deceiving victims into performing unauthorized actions), business email compromise (i.e., high-ranking official deepfakes as a tool for illicit data sharing or fund transfers), authentication bypass (i.e., audio or visual deepfake deceiving voice or face authentication systems, allowing unauthorized access to secure systems or sensitive information), and insider threats (i.e., deepfakes impersonating employees, gaining unauthorized access, or manipulating others within an organization for malicious purposes such as espionage).

Current cybersecurity defense mechanisms should be revisited in order to effectively counter deepfake threats. Defending against deepfake cybersecurity attacks requires a multi-faceted strategy involving technology, education, and vigilance. Some of the key countermeasures are behavioral analytics (i.e., monitoring user behavior and detecting anomalies that may indicate a deepfake), cutting-edge security solutions (i.e., advanced email filtering and endpoint protection solutions to identify and block phishing attempts using deepfake technology), multi-factor authentication (i.e., to add an extra layer of security for identity verification), robust anti-spoofing (i.e., audio, visual, and audiovisual mechanisms to detect deepfake attempts on user authentication systems), secure communication channels (i.e., encrypted messaging platforms for sensitive and confidential information),

deepfake detection tools usable in core cybersecurity networks, watermarking and digital signatures (i.e., incorporating digital watermarks and/or signatures in multimedia content to verify its authenticity), continuous monitoring (i.e., proactive approach for network traffic and user activities to detect deepfakes in real-time), employee training and awareness (i.e., novel ways to educate employees on deepfake awareness, associated risks, and threat recognition), AI-powered threat intelligence platforms (i.e., tools analyzing large datasets to identify emerging deepfake threats and provide actionable insights to enhance cybersecurity defenses), intrusion detection and prevention systems (i.e., network and/or system monitoring systems to detect deepfake activities and automatically responding to potential threats), and blockchain (i.e., blockchain-based timestamping of multimedia content to establish the authenticity).

4.31. Reproducible Research

The research and professional communities should advocate the reproducible deepfake research trend [546]. This can be accomplished by enriching large public databases with comprehensive human assigned deepfakability scores and reasons, offering clear and accessible experimental setups, and providing open-source deepfake and audio/face manipulation codes and tools. This will ensure an accurate assessment of deepfake technology progress while averting the overestimation of deepfake algorithmic strengths.

4.32. Forensics Specialists with Limited Experience in Courtrooms

The use of audio, visual, and multimedia samples as legal evidence relies on digital media forensics experts. Recent advancements in digital content manipulation pose a greater challenge for detection among digital forensics experts with training in law enforcement and/or computer science [547,548]. Digital media forensics professionals must remain vigilant regarding the ease with which multimedia content can be manipulated, as such false evidences may lead to wrongful convictions. The complexity of deepfakes in courtrooms is compounded by the silent witness theory (i.e., audio, photos, and videos are considered to speak for themselves as legal evidence). Nonetheless, an empirical study has shown that attackers can remotely infiltrate body camera devices, extract footage, make alterations, or selectively delete portions they wish to conceal from law enforcement, and seamlessly re-upload the manipulated content without leaving any identifiable trace of tampering [549]. Digital media forensic results, vital for court use, require meticulous validation; however, AI-based manipulation detectors, while accurate, lack explainability (i.e., black-box models). The straightforward incorporation of AI-based detectors into forensic applications is hindered, as digital forensics experts and computer professionals lack the necessary knowledge of AI algorithms and struggle to explain the results effectively. New courses specifically tailored for digital media forensics experts should be developed. New forensic models that combine cyber-forensic and incident response approaches would empower forensics experts to conduct thorough and legally sound investigations. Future research directions should concentrate on deepfake forensic tools that are explainable, accurate, and meaningful to meet the stringent standards for admissibility in legal proceedings.

4.33. National and International Competitions and Challenges for Deepfake Technology

In the spirit of fostering innovation and technological progress, public competitions and challenges have been launched, inviting individuals or teams from the general public, academia, and industry to showcase their deepfake skills within a competitive framework, e.g., the Open Media Forensics Challenge [550] by NIST, Deepfake Detection Challenge [85], ADD [328], and ASVSpoof [320]. These competitions and challenges have mainly concentrated on the development of deepfake detectors rather than generators or a combination of both. Moreover, these events are held biannually and less frequently. More frequent national and international deepfake competitions and challenges must be hosted, taking

into account factors such as diverse data quality, sample variability, model complexity, human detectors, evaluation metrics, and ethical and social implications.

4.34. Deepfake Education for the General Population

Deepfakes are rapidly on the rise, posing a growing threat to individuals in online spaces, including audio and video calls. A global study regarding public deepfake media knowledge was conducted by Iproov in 2022 and found that around only 29% of the population understands deepfakes [551]. Moreover, tech-savvy individuals, despite their social media literacy, often feel a confidence dip when exposed to deepfake detection outcomes. Hence, it is paramount to educate the general public about the perils of deepfakes and equip them with the skills to critically evaluate digital content and detect early indicators of a deepfake attacks. By enhancing individuals' ability to identify deepfakes, we can effectively minimize the impact of deceptive contents. The realm of deepfake technology literacy is still in its early stages of development. Researchers and professionals must dedicate additional efforts to formulate effective training and educational frameworks for the general public. Furthermore, they should identify and assess both existing and novel educational strategies. Multilingual online tutorials, videos, and presentations could prove highly beneficial on a global scale.

4.35. Open-Source Intelligence (OSINT) Techniques against Deepfake

The deepfake OSINT approaches aim at devising and sharing open-source tools to identify disinformation-related contents and deepfakes [552]. The widely used approach is reverse image search (e.g., FotoForensics [553], InVID [554], and WeVerify [555]), which aids a user in verifying the authenticity of a questionable image or video. Encyclopedic OSINT tools against deepfakes should be designed to provide superior accuracy and quality of obtained search results. These tools should include features like audio, visual, and multimedia deepfake or audio/face tampering detection, reverse or audio/image/video search, metadata analysis, noise analysis, voice clone detection, extensive datasets, and user-friendly interfaces. There is a shortage of OSINT tools for audio, 3D, virtual reality, and mixed reality deepfakes, along with limited large-scale deepfake OSINT databases.

4.36. Interdisciplinary Research

The research community must prioritize fostering interdisciplinary basic science to cultivate resilient, authentic, reliable, and universally applicable techniques for effectively addressing challenges posed by deepfakes, fake news, disinformation, and misinformation [556–558]. Any such efforts will advance the state-of-the-art fake content and information detection technologies.

5. Conclusions

The substantive evolutions in AI and ML are unlocking new possibilities across sectors like finance, healthcare, manufacturing, entertainment, information dissemination, and storage and retrieval. Deepfakes leverage AI and ML advancements to digitally manipulate or create convincing and highly realistic audio, visual, or audiovisual face samples. Deepfakes have the potential to deceive speech/facial recognition technologies; deliver malicious payloads in state-sponsored cyber-warfare; and propagate misinformation, disinformation, and fake news, thereby eroding the privacy, security, trustworthiness of online contents and democratic stability. Researchers and practitioners strive to enhance deepfake detection, yet a fierce competition persists between generator and detector advancements. Thus, this paper first presents an extensive overview of current audio, image, and video deepfake and audio/face manipulation databases with detailed insights into their characteristics and nuances. Next, the paper extensively explores the open challenges and promising research directions in audio, image, and video deepfake generation and mitigation. Concerted interdisciplinary efforts are essential for making significant headway in deepfake technology, and this article offers a valuable reference point for creating cutting-edge deepfake

generation and detection algorithms. In summary, this article strives to augment existing survey studies and serve as a catalyst to inspire newcomers; researchers; practitioners; engineers; information-, political-, and social-scientists; policymakers; and multimedia forensic investigators to consider deepfakes as a focal point in their scholarly pursuits.

Author Contributions: Conceptualization, Z.A.; methodology, Z.A.; formal analysis, Z.A., T.L.P., and V.S.A.; investigation, Z.A., T.L.P., and V.S.A.; writing—original draft preparation, Z.A., T.L.P., and V.S.A.; writing—review and editing, Z.A.; supervision, Z.A.; project administration, Z.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data sharing not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Spector, N. Available online: https://www.nbcnews.com/business/consumer/so-it-s-fine-if-you-edit-your-selfies-not-n766186 (accessed on 16 December 2023).
2. Akhtar, Z. Deepfakes Generation and Detection: A Short Survey. *J. Imaging* **2023**, *9*, 18. [CrossRef] [PubMed]
3. Thomson, T.J.; Angus, D.; Dootson, P. Available online: https://theconversation.com/3-2-billion-images-and-720-000-hours-of-video-are-shared-online-daily-canyou-sort-real-from-fake-148630 (accessed on 16 December 2023).
4. Akhtar, Z.; Dasgupta, D. A comparative evaluation of local feature descriptors for deepfakes detection. In Proceedings of the 2019 IEEE International Symposium on Technologies for Homeland Security (HST), Woburn, WA, USA, 5 November 2019; pp. 1–19.
5. Juefei-Xu, F.; Wang, R.; Huang, Y.; Guo, Q.; Ma, L.; Liu, Y. Countering malicious deepfakes: Survey, battleground, and horizon. *Int. J. Comput. Vis.* **2022**, *130*, 1678–1734. [CrossRef]
6. Seow, J.W.; Lim, M.K.; Phan, R.C.; Liu, J.K. A comprehensive overview of Deepfake: Generation, detection, datasets, and opportunities. *Neurocomputing* **2022**, *513*, 351–371. [CrossRef]
7. Shahzad, H.F.; Rustam, F.; Flores, E.S.; Luís Vidal Mazón, J.; de la Torre Diez, I.; Ashraf, I. A Review of Image Processing Techniques for Deepfakes. *Sensors* **2022**, *22*, 4556. [CrossRef]
8. Mirsky, Y.; Lee, W. The creation and detection of deepfakes: A survey. *ACM Comput. Surv.* **2021**, *54*, 1–41. [CrossRef]
9. Akhtar, Z.; Dasgupta, D.; Banerjee, B. Face Authenticity: An Overview of Face Manipulation Generation, Detection and Recognition. In Proceedings of the International Conference on Communication and Information Processing (ICCIP), Pune, India, 17–18 May 2019; pp. 1–8.
10. FaceApp Technology Limited. Available online: https://www.faceapp.com/ (accessed on 16 December 2023).
11. Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; Aila, T. Analyzing and improving the image quality of stylegan. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 8110–8119.
12. Almutairi, Z.; Elgibreen, H. A review of modern audio deepfake detection methods: Challenges and future directions. *Algorithms* **2022**, *15*, 155. [CrossRef]
13. Mathew, J.J.; Ahsan, R.; Furukawa, S.; Kumar, J.G.; Pallan, H.; Padda, A.S.; Adamski, S.; Reddiboina, M.; Pankajakshan, A. Towards the Development of a Real-Time Deepfake Audio Detection System in Communication Platforms. *arXiv* **2024**, arXiv:2403.11778.
14. Yi, J.; Wang, C.; Tao, J.; Zhang, X.; Zhang, C.Y.; Zhao, Y. Audio Deepfake Detection: A Survey. *arXiv* **2023**, arXiv:2308.14970.
15. Khanjani, Z.; Gabrielle, W.; Janeja, V.P. How deep are the fakes? focusing on audio deepfake: A survey. *arXiv* **2021**, arXiv:2111.14203.
16. Altuncu, E.; Franqueira, V.N.; Li, S. Deepfake: Definitions, Performance Metrics and Standards, Datasets and Benchmarks, and a Meta-Review. *arXiv* **2022**, arXiv:2208.10913.
17. Liz-López, H.; Keita, M.; Taleb-Ahmed, A.; Hadid, A. Huertas-Tato J, Camacho D. Generation and detection of manipulated multimodal audiovisual content: Advances, trends and open challenges. *Inf. Fusion* **2024**, *103*, 102103. [CrossRef]
18. Resemble AI. Available online: https://www.resemble.ai/cloned/ (accessed on 16 December 2023).
19. Xu, Z.; Hong, Z.; Ding, C.; Zhu, Z.; Han, J.; Liu, J.; Ding, E. MobileFaceSwap: A Lightweight Framework for Video Face Swapping. *Aaai Conf. Artif. Intell.* **2022**, *36*, 2973–2981. [CrossRef]
20. Jia, Y.; Zhang, Y.; Weiss, R.; Wang, Q.; Shen, J.; Ren, F.; Nguyen, P.; Pang, R.; Lopez Moreno, I.; Wu, Y. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 2 December 2018; Volume 31, pp. 1–11.

21. Damiani, J. A Voice DeepfakeWas Used to Scam a CEO Out of $243,000. 2019. Available online: https://www.forbes.com/sites/jessedamiani/2019/09/03/a-voice-deepfake-was-used-to-scam-a-ceo-out-of-243000/?sh=173f55a52241 (accessed on 16 December 2023).
22. Shu, C.; Wu, H.; Zhou, H.; Liu, J.; Hong, Z.; Ding, C.; Han, J.; Liu, J.; Ding, E.; Wang, J. Few-Shot Head Swapping in the Wild. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 10789–10798.
23. Wang, J.; Wu, Z.; Ouyang, W.; Han, X.; Chen, J.; Jiang, Y.; Li, S. M2TR: Multi-modal multi-scale transformers for deepfake detection. In Proceedings of the International Conference on Multimedia Retrieval, Newark, NJ, USA, 27–30 June 2022; pp. 615–623.
24. Agarwal, M.; Mukhopadhyay, R.; Namboodiri, V.; Jawahar, C. Audio-visual face reenactment. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–7 January 2023; pp. 5178–5187.
25. Cozzolino, D.; Pianese, A.; Nießner, M.; Verdoliva, L. Audio-visual person-of-interest deepfake detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 943–952.
26. Patashnik, O.; Wu, Z.; Shechtman, E.; Cohen-Or, D.; Lischinski, D. Styleclip: Text-driven manipulation of stylegan imagery. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 10–17 March 2021; pp. 2085–2094.
27. Asnani, V.; Yin, X.; Hassner, T.; Liu, X. Malp: Manipulation localization using a proactive scheme. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 12343–12352.
28. Li, Z.; Min, M.; Li, K.; Xu, C. StyleT2I: Toward Compositional and High-Fidelity Text-to-Image Synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 18197–18207.
29. Tan, C.; Zhao, Y.; Wei, S.; Gu, G.; Wei, Y. Learning on Gradients: Generalized Artifacts Representation for GAN-Generated Images Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 12105–12114.
30. Luong, H.T.; Yamagishi, J. Nautilus: A versatile voice cloning system. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2020**, *28*, 2967–2981. [CrossRef]
31. Kulangareth, N.V.; Kaufman, J.; Oreskovic, J.; Fossat, Y. Investigation of Deepfake Voice Detection Using Speech Pause Patterns: Algorithm Development and Validation. *Jmir Biomed. Eng.* **2024**, *9*, e56245. [CrossRef] [PubMed]
32. Oord, A.; Li, Y.; Babuschkin, I.; Simonyan, K.; Vinyals, O.; Kavukcuoglu, K.; Driessche, G.; Lockhart, E.; Cobo, L.; Stimberg, F.; et al. Parallel wavenet: Fast high-fidelity speech synthesis. In Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 10–15 June 2018; pp. 3918–3926.
33. Rahman, M.H.; Graciarena, M.; Castan, D.; Cobo-Kroenke, C.; McLaren, M.; Lawson, A. Detecting synthetic speech manipulation in real audio recordings. In Proceedings of the 2022 IEEE International Workshop on Information Forensics and Security (WIFS), Shanghai, China, 12–16 December 2022; pp. 1–6.
34. Wang, R.; Ding, Y.; Li, L.; Fan, C. One-shot voice conversion using star-gan. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Virtual, 4–8 May 2020; pp. 7729–7733.
35. Lo, C.C.; Fu, S.W.; Huang, W.C.; Wang, X.; Yamagishi, J.; Tsao, Y.; Wang, H.M. Mosnet: Deep learning based objective assessment for voice conversion. *arXiv* **2019**, arXiv:1904.08352.
36. Choi, W.; Kim, M.; Martínez Ramírez, M.A.; Chung, J.; Jung, S. Amss-net: Audio manipulation on user-specified sources with textual queries. In Proceedings of the 29th ACM International Conference on Multimedia, Chengdu, China, 20–24 October 2021; pp. 1775–1783.
37. Zhao, Y.; Yi, J.; Tao, J.; Wang, C.; Zhang, X.; Dong, Y. EmoFake: An initial dataset for emotion fake audio detection. *arXiv* **2022**, arXiv:2211.05363.
38. Jia, Y.; Ramanovich, M.T.; Remez, T.; Pomerantz, R. Translatotron 2: High-quality direct speech-to-speech translation with voice preservation. In Proceedings of the International Conference on Machine Learning, Baltimore, MD, USA, 17–23 July 2022; pp. 10120–10134.
39. Kuo, H.K.; Kislal, E.E.; Mangu, L.; Soltau, H.; Beran, T. Out-of-vocabulary word detection in a speech-to-speech translation system. In Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 4–9 May 2014; pp. 7108–7112.
40. Yi, J.; Bai, Y.; Tao, J.; Ma, H.; Tian, Z.; Wang, C.; Wang, T.; Fu, R. Half-truth: A partially fake audio detection dataset. *arXiv* **2021**, arXiv:2104.03617.
41. Wu, H.; Kuo, H.C.; Zheng, N.; Hung, K.H.; Lee, H.Y.; Tsao, Y.; Wang, H.M.; Meng, H. Partially fake audio detection by self-attention-based fake span discovery. In Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Virtual, 22–27 May 2022; pp. 9236–9240.
42. Dong, S.; Wang, J.; Ji, R.; Liang, J.; Fan, H.; Ge, Z. Implicit Identity Leakage: The Stumbling Block to Improving Deepfake Detection Generalization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 3994–4004.
43. Mehta, V.; Gupta, P.; Subramanian, R.; Dhall, A. FakeBuster: A DeepFakes detection tool for video conferencing scenarios. In Proceedings of the International Conference on Intelligent User Interfaces-Companion, College Station, TX, USA, 13–17 April 2021; pp. 61–63.

44. Hussain, S.; Neekhara, P.; Jere, M.; Koushanfar, F.; McAuley, J. Adversarial deepfakes: Evaluating vulnerability of deepfake detectors to adversarial examples. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Virtual, 5–9 January 2021; pp. 3348–3357.
45. Masood, M.; Nawaz, M.; Malik, K.M.; Javed, A.; Irtaza, A.; Malik, H. Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward. *Appl. Intell.* **2023**, *53*, 3974–4026. [CrossRef]
46. Naitali, A.; Ridouani, M.; Salahdine, F.; Kaabouch, N. Deepfake Attacks: Generation, Detection, Datasets, Challenges, and Research Directions. *Computers* **2023**, *12*, 216. [CrossRef]
47. Shaaban, O.A.; Yildirim, R.; Alguttar, A.A. Audio Deepfake Approaches. *IEEE Access* **2023**, *11*, 132652–132682. [CrossRef]
48. Dagar, D.; Vishwakarma, D.K. A literature review and perspectives in deepfakes: Generation, detection, and applications. *Int. J. Multimed. Inf. Retr.* **2022**, *11*, 219–289. [CrossRef]
49. De Carvalho, T.J.; Riess, C.; Angelopoulou, E.; Pedrini, H.; de Rezende Rocha, A. Exposing digital image forgeries by illumination color classification. *IEEE Trans. Inf. Forensics Secur. (TIFS)* **2013**, *8*, 1182–1194. [CrossRef]
50. Karras, T.; Aila, T.; Laine, S.; Lehtinen, J. Progressive growing of gans for improved quality, stability, and variation. *arXiv* **2017**, arXiv:1710.10196.
51. Korshunov, P.; Marcel, S. Deepfakes: A new threat to face recognition? assessment and detection. *arXiv* **2018**, arXiv:1812.08685.
52. Faceswap-GAN. Available online: https://github.com/shaoanlu/faceswap-GAN (accessed on 7 January 2023).
53. CelebA-HQ Download Link. Available online: https://drive.google.com/drive/folders/0B4qLcYyJmiz0TXY1NG02bzZVRGs?resourcekey=0-arAVTUfW9KRhN-irJchVKQ (accessed on 17 April 2024).
54. DeepfakeTIMIT Download Link. Available online: https://zenodo.org/records/4068245 (accessed on 24 April 2024).
55. Li, Y.; Chang, M.C.; Lyu, S. In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In Proceedings of the 2018 IEEE International Workshop on Information Forensics and Security (WIFS), Hong Kong, 11–13 December 2018; pp. 1–7.
56. EBV (Eye Blinking Video Dataset) Download Link. Available online: https://1drv.ms/u/s!As4tun0sWfKsgdVcYJ-nn0bw0kdjzw?e=DAcGfb (accessed on 24 April 2024).
57. Rössler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; Nießner, M. Faceforensics: A large-scale video dataset for forgery detection in human faces. *arXiv* **2018**, arXiv:1803.09179.
58. Sohan, M.F.; Solaiman, M.; Hasan, M.A. A survey on deepfake video detection datasets. *Indones. J. Electr. Eng. Comput. Sci.* **2023**, *32*, 1168–1176. [CrossRef]
59. Thies, J.; Zollhofer, M.; Stamminger, M.; Theobalt, C.; Nießner, M. Face2face: Real-time face capture and reenactment of rgb videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2387–2395.
60. Faceforensics Link. Available online: https://github.com/ondyari/FaceForensics/tree/original (accessed on 31 December 2023).
61. Khodabakhsh, A.; Ramachandra, R.; Raja, K.; Wasnik, P.; Busch, C. Fake face detection methods: Can they be generalized? In Proceedings of the 2018 International Conference of the Biometrics Special Interest Group (BIOSIG), Darmstadt, Germany, 26–28 September 2018; pp. 1–6.
62. FFW Dataset Download. Available online: https://github.com/AliKhoda/FFW/blob/main/download.py (accessed on 17 April 2024).
63. Güera, D.; Delp, E.J. Deepfake video detection using recurrent neural networks. In Proceedings of the 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Auckland, New Zealand, 27–30 November 2018; pp. 1–6.
64. Yang, X.; Li, Y.; Lyu, S. Exposing deep fakes using inconsistent head poses. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12 May 2019; pp. 8261–8265.
65. Cai, Z.; Ghosh, S.; Adatia, A.P.; Hayat, M.; Dhall, A.; Stefanov, K. AV-Deepfake1M: A Large-Scale LLM-Driven Audio-Visual Deepfake Dataset. *arXiv* **2023**, arXiv:2311.15308.
66. Waseem, S.; Abu-Bakar, S.R.; Ahmed, B.A.; Omar, Z.; Eisa, T.A.; Dalam, M.E. DeepFake on Face and Expression Swap: A Review. *IEEE Access* **2023**, *11*, 117865–117906. [CrossRef]
67. UADFV Dataset Download Link. Available online: https://drive.google.com/file/d/17d-0K2UblFldBmjTUk3_nASK8MhhiSHa/view (accessed on 24 April 2024).
68. Li, Y.; Yang, X.; Sun, P.; Qi, H.; Lyu, S. Celeb-DF: A large-scale challenging dataset for deepfake forensics. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 3207–3216.
69. Celeb-DF Download Link. Available online: https://drive.google.com/file/d/1iLx76wsbi9itnkxSqz9BVBl4ZvnbIazj/view (accessed on 24 April 2024).
70. Dufour, N.; Gully, A.; Karlsson, P.; Vorbyov, A.V.; Leung, T.; Childs, J.; Bregler, C. DeepFakes Detection Dataset by Google & JigSaw. 2019. Available online: https://research.google/blog/contributing-data-to-deepfake-detection-research/ (accessed on 4 July 2024).
71. FaceForensics++ Link. Available online: https://github.com/ondyari/FaceForensics (accessed on 8 January 2024).
72. Mahfoudi, G.; Tajini, B.; Retraint, F.; Morain-Nicolier, F.; Dugelay, J.L.; Marc, P.I. DEFACTO: Image and face manipulation dataset. In Proceedings of the 2019 27th European Signal Processing Conference (EUSIPCO), A Coruna, Spain, 2–6 September 2019; pp. 1–5.
73. DEFACTO Dataset Download Link. Available online: https://www.kaggle.com/defactodataset/datasets (accessed on 24 April 2024)

74. Dang, H.; Liu, F.; Stehouwer, J.; Liu, X.; Jain, A.K. On the detection of digital face manipulation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition, Virtual, 14–19 June 2020; pp. 5781–5790.
75. DFFD Dataset Download Link. Available online: https://cvlab.cse.msu.edu/dffd-dataset.html (accessed on 24 April 2024).
76. Rössler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; Nießner, M. Faceforensics++: Learning to detect manipulated facial images. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Long Beach, CA, USA, 15–20 June 2019; pp. 1–11.
77. Sultan, D.A.; Ibrahim, L.M. A Comprehensive Survey on Deepfake Detection Techniques. *Int. J. Intell. Syst. Appl. Eng.* **2022**, *10*, 189–202.
78. Karras, T.; Laine, S.; Aila, T. A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 4401–4410.
79. FFHQ Dataset. Available online: https://github.com/NVlabs/ffhq-dataset (accessed on 1 January 2024).
80. Whichfaceisreal. Available online: https://www.whichfaceisreal.com/ (accessed on 1 January 2024).
81. Jiang, L.; Li, R.; Wu, W.; Qian, C.; Loy, C.C. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 14–19 June 2020; pp. 2889–2898.
82. Jiang, L.; Guo, Z.; Wu, W.; Liu, Z.; Liu, Z.; Loy, C.C.; Yang, S.; Xiong, Y.; Xia, W.; Chen, B.; et al. DeeperForensics Challenge 2020 on real-world face forgery detection: Methods and results. *arXiv* **2021**, arXiv:2102.09471.
83. Available online: https://github.com/EndlessSora/DeeperForensics-1.0/tree/master/dataset#download (accessed on 31 December 2023).
84. DeeperForensics Dataset Download Link. Available online: https://drive.google.com/drive/folders/1s3KwYyTIXT78VzkRazn9QDPuNh18TWe- (accessed on 18 April 2024)
85. Dolhansky, B.; Bitton, J.; Pflaum, B.; Lu, J.; Howes, R.; Wang, M.; Ferrer, C.C. The deepfake detection challenge (dfdc) dataset. *arXiv* **2020**, arXiv:2006.07397.
86. The Deepfake Detection Challenge (dfdc) Dataset Download Link. Available online: https://www.kaggle.com/competitions/deepfake-detection-challenge/data (accessed on 24 April 2024).
87. Akhtar, Z.; Mouree, M.R.; Dasgupta, D. Utility of deep learning features for facial attributes manipulation detection. In Proceedings of the 2020 IEEE International Conference on Humanized Computing and Communication with Artificial Intelligence (HCCAI), Irvine, CA, USA, 21–23 September 2020; pp. 55–60.
88. Li, L.; Bao, J.; Yang, H.; Chen, D.; Wen, F. Faceshifter: Towards high fidelity and occlusion aware face swapping. *arXiv* **2019**, arXiv:1912.13457.
89. Gupta, P.; Chugh, K.; Dhall, A.; Subramanian, R. The eyes know it: Fakeet-an eye-tracking database to understand deepfake perception. In Proceedings of the 2020 International Conference on Multimodal Interaction, Utrecht, The Netherlands, 21 October 2020; pp. 519–527.
90. FakeET Dataset Download Link. Available online: https://drive.google.com/drive/folders/1DpDIbjRTn3rTVdc5PU9uprRdLfmRgr-8?usp=sharing_eil_m&ts=655e5535 (accessed on 18 April 2024).
91. Zhou, T.; Wang, W.; Liang, Z.; Shen, J. Face forensics in the wild. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 5778–5788.
92. FFIW10K Train Dataset Download Link. Available online: https://drive.google.com/file/d/1-Ha_A9yRFS0dACrv-L156Kfy_yaPn980/view?usp=sharing (accessed on 18 April 2024).
93. FFIW10K Test Dataset Download Link. Available online: https://drive.google.com/file/d/1ydNrV_LK3Ep6i3_WPsUo0_aQan4kDUbQ/view?usp=sharing (accessed on 18 April 2024).
94. Neves, J.C.; Tolosana, R.; Vera-Rodriguez, R.; Lopes, V.; Proença, H.; Fierrez, J. Ganprintr: Improved fakes and evaluation of the state of the art in face manipulation detection. *IEEE J. Sel. Top. Signal Process.* **2020**, 14, 1038–1048. [CrossRef]
95. Available online: https://github.com/socialabubi/iFakeFaceDB (accessed on 1 January 2024).
96. iFakeFaceDB Dataset Download Link. Available online: http://socia-lab.di.ubi.pt/~jcneves/iFakeFaceDB.zip (accessed on 18 April 2024).
97. Nam, S.; Oh, S.W.; Kang, J.Y.; Shin, C.H.; Jo, Y.; Kim, Y.H.; Kim, K.; Shim, M.; Lee, S.; Kim, Y.; et al. Real and Fake Face Detection, Version 1. 2019. Available online: https://www.kaggle.com/datasets/ciplab/real-and-fake-face-detection (accessed on 1 January 2024).
98. Oliver, M.M.; Amengual Alcover, E. UIBVFED: Virtual facial expression dataset. *PLoS ONE* **2020**, *15*, e0231266. [CrossRef]
99. UIBVFED Dataset Download Link. Available online: https://ugivia.uib.es/uibvfed/ (accessed on 18 April 2024).
100. Zi, B.; Chang, M.; Chen, J.; Ma, X.; Jiang, Y.G. Wilddeepfake: A challenging real-world dataset for deepfake detection. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 2382–2390.
101. WildDeepFake Dataset Download Link. Available online: https://drive.google.com/drive/folders/1Cb_OqksBU3x7HFIo8EvDTigU6IjM7tmp (accessed on 18 April 2024).
102. Kukanov, I.; Karttunen, J.; Sillanpää, H.; Hautamäki, V. Cost sensitive optimization of deepfake detector. In Proceedings of the 2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Auckland, New Zealand, 7–10 December 2020; pp. 1300–1303.
103. Huang, J.; Wang, X.; Du, B.; Du, P.; Xu, C. DeepFake MNIST+: A DeepFake facial animation dataset. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021; pp. 1973–1982.

104. Deepfake MNIST+ Dataset Download Link. Available online: https://1fichier.com/?do5lezggwcnpg49m28wh (accessed on 18 April 2024).
105. Siarohin, A.; Lathuilière, S.; Tulyakov, S.; Ricci, E.; Sebe, N. First order motion model for image animation. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; pp. 1–11.
106. Alamayreh, O.; Barni, M. Detection of gan-synthesized street videos. In Proceedings of the 2021 IEEE 29th European Signal Processing Conference (EUSIPCO), Dublin, Ireland, 23 August 2021; pp. 811–815.
107. DeepStreets Dataset Download Link. Available online: http://clem.dii.unisi.it/~vipp/datasets.html (accessed on 18 April 2024).
108. Peng, B.; Fan, H.; Wang, W.; Dong, J.; Li, Y.; Lyu, S.; Li, Q.; Sun, Z.; Chen, H.; Chen, B.; et al. DFGC 2021: A deepfake game competition. In Proceedings of the 2021 IEEE International Joint Conference on Biometrics (IJCB), Shenzhen, China, 4–7 August 2021; pp. 1–8.
109. DFGC-21 Dataset Download Link. Available online: https://drive.google.com/drive/folders/1SD4L3R0XCZnr-LnZy5G9Vsho9BpIYe6Z (accessed on 18 April 2024).
110. Jain, A.; Korshunov, P.; Marcel, S. Improving generalization of deepfake detection by training for attribution. In Proceedings of the 2021 IEEE 23rd International Workshop on Multimedia Signal Processing (MMSP), Tampere, Finland, 6 October 2021; pp. 1–6.
111. DF-Mobio Dataset Download Link. Available online: https://zenodo.org/records/5769057 (accessed on 3 July 2024).
112. Pu, J.; Mangaokar, N.; Kelly, L.; Bhattacharya, P.; Sundaram, K.; Javed, M.; Wang, B.; Viswanath, B. Deepfake Videos in the Wild: Analysis and Detection. In Proceedings of the Web Conference 2021, Ljubljana, Slovenia, 19–23 April 2021; pp. 981–992.
113. DF-W Dataset Download Link. Available online: https://drive.google.com/drive/folders/1j6OaWK-4ZQquV7Y3xWD5zYa0TizRa4B6 (accessed on 18 April 2024).
114. Wood, E.; Baltrušaitis, T.; Hewitt, C.; Dziadzio, S.; Cashman, T.J.; Shotton, J. Fake it till you make it: Face analysis in the wild using synthetic data alone. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021; pp. 3681–3691.
115. Available online: https://github.com/microsoft/FaceSynthetics (accessed on 1 January 2024).
116. Khalid, H.;Tariq, S.; Kim, M.; Woo, S.S. FakeAVCeleb: A novel audio-video multimodal deepfake dataset. *arXiv* **2021**, arXiv:2108.05080.
117. Fakeavceleb Dataset Link. Available online: https://sites.google.com/view/fakeavcelebdash-lab/download?authuser=0 (accessed on 1 January 2024).
118. FakeAvCeleb Dataset Download Link. Available online: https://drive.google.com/file/d/1x0h3mhmfqWErN9xAq7mUfn6EcbUPIDMa/view?ts=652e15e1 (accessed on 18 April 2024).
119. He, Y.; Gan, B.; Chen, S.; Zhou, Y.; Yin, G.; Song, L.; Sheng, L.; Shao, J.; Liu, Z. Forgerynet: A versatile benchmark for comprehensive forgery analysis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 4360–4369.
120. ForgeryNet Dataset Download Link. Available online: https://opendatalab.com/OpenDataLab/ForgeryNet/tree/main (accessed on 18 April 2024).
121. Wang, Y.; Chen, X.; Zhu, J.; Chu, W.; Tai, Y.; Wang, C.; Li, J.; Wu, Y.; Huang, F.; Ji, R. Hififace: 3d shape and semantic prior guided high fidelity face swapping. *arXiv* **2021**, arXiv:2106.09965.
122. HifiFace Dataset Link. Available online: https://johann.wang/HifiFace/ (accessed on 8 January 2024).
123. HiFiFace Dataset Download Link. Available online: https://drive.google.com/file/d/1tZitaNRDaIDK1MPOaQJJn5CivnEIKMnB/view (accessed on 18 April 2024).
124. Kwon, P.; You, J.; Nam, G.; Park, S.; Chae, G. KoDF: A large-scale korean deepfake detection dataset. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021; pp. 10744–10753.
125. KODF Dataset Download Link. Available online: https://deepbrainai-research.github.io/kodf/ (accessed on 18 April 2024).
126. Le, T.N.; Nguyen, H.H.; Yamagishi, J.; Echizen, I. Openforensics: Large-scale challenging dataset for multi-face forgery detection and segmentation in-the-wild. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021; pp. 10117–10127.
127. OpenForensics Dataset Download Link. Available online: https://zenodo.org/records/5528418 (accessed on 18 April 2024).
128. Lago, F.; Pasquini, C.; Böhme, R.; Dumont, H.; Goffaux, V.; Boato, G. More real than real: A study on human visual perception of synthetic faces [applications corner]. *IEEE Signal Process. Mag.* **2021**, *39*, 109–116. [CrossRef]
129. Perception Synthetic Face Dataset Download Link. Available online: https://drive.google.com/drive/folders/1d7JhLnXu7r5fm2uZs4EyjGLrkwSgFdYB (accessed on 18 April 2024).
130. Fox, G.; Liu, W.; Kim, H.; Seidel, H.-P.; Elgharib, M.; Theobalt, C. Video-Forensics-HQ: Detecting High-quality Manipulated Face Videos. In Proceedings of the IEEE International Conference on Multimedia and Expo (ICME 2021), Shenzhen, China (Virtual), 5–9 July 2021.
131. VideoForensicsHQ Dataset Download Link. Available online: https://nextcloud.mpi-klsb.mpg.de/index.php/s/EW9bCwCPisfFpww (accessed on 18 April 2024).
132. Mittal, T.; Sinha, R.; Swaminathan, V.; Collomosse, J.; Manocha, D. Video Manipulations Beyond Faces: A Dataset with Human-Machine Analysis. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–7 January 2023; pp. 643–652.

133. Lin, M.; Cao, G.; Lou, Z. Spatio-temporal Co-attention Fusion Network for Video Splicing Localization. *arXiv* **2023**, arXiv:2309.09482.
134. VideoSham Dataset Download Link. Available online: https://github.com/adobe-research/VideoSham-dataset (accessed on 18 April 2024).
135. Yavuzkiliç, S.; Sengur, A.; Akhtar, Z.; Siddique, K. Spotting DeepFakes and Face Manipulations by Fusing Features from Multi-Stream CNNs Models. *Symmetry* **2021**, *13*, 1352. [CrossRef]
136. Li, C.; Huang, Z.; Paudel, D.P.; Wang, Y.; Shahbazi, M.; Hong, X.; Van Gool, L. A continual deepfake detection benchmark: Dataset, methods, and essentials. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–7 January 2023; pp. 1339–1349.
137. CDDB Dataset Download Link. Available online: https://drive.google.com/file/d/1NgB8ytBMFBFwyXJQvdVT_yek1EaaEHrg/view (accessed on 18 April 2024).
138. Zhu, H.; Wu, W.; Zhu, W.; Jiang, L.; Tang, S.; Zhang, L.; Liu, Z.; Loy, C.C. CelebV-HQ: A large-scale video facial attributes dataset. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23 October 2022; pp. 650–667.
139. CELEBV-HQ Dataset Download Link. Available online: https://pan.baidu.com/s/1TGzOwUcXsRw72l4gaWre_w?pwd=pg71#list/path=%2F (accessed on 18 April 2024).
140. Narayan, K.; Agarwal, H.; Thakral, K.; Mittal, S.; Vatsa, M.; Singh, R. Deephy: On deepfake phylogeny. In Proceedings of the IEEE International Joint Conference on Biometrics (IJCB), Abu Dhabi, UAE, 10 October 2022; pp. 1–10.
141. DeePhy Dataset Download Link. Available online: https://drive.google.com/file/d/1xbFOITBiYe74Oo5_5jysYpprcc3iLFDW/view (accessed on 18 April 2024).
142. Jia, S.; Li, X.; Lyu, S. Model attribution of face-swap deepfake videos. In Proceedings of the 2022 IEEE International Conference on Image Processing (ICIP), Bordeaux, France, 16 October 2022; pp. 2356–2360.
143. DFDM Dataset Download Link. Available online: https://drive.google.com/drive/folders/1aXxeMdA2qwjDytyIgr4CBMVy4pAWizdX (accessed on 18 April 2024).
144. Alamayreh, O.; Fascella, C.; Mandelli, S.; Tondi, B.; Bestagini, P.; Barni, M. Just Dance: Detection of human body reenactment fake videos. *EURASIP J. Image Video Process.* 2022, *under review*. [CrossRef]
145. FakeDance Dataset Download Link. Available online: https://drive.google.com/drive/folders/1IoMA0kEx1IJJVEK0XRR4uPoZtACm6FwC (accessed on 18 April 2024).
146. Li, G.; Zhao, X.; Cao, Y.; Pei, P.; Li, J.; Zhang, Z. FMFCC-V: An Asian Large-Scale Challenging Dataset for DeepFake Detection. In Proceedings of the 2022 ACM Workshop on Information Hiding and Multimedia Security, Santa Barbara, CA, USA, 27–28 June 2022; pp. 7–18.
147. FMFCC-V Dataset Download Link. Available online: https://pan.baidu.com/s/1wF87JgSbX_buqsc4eX-2FQ#list/path=%2F (accessed on 18 April 2024).
148. Nadimpalli, A.V.; Rattani, A. GBDF: Gender balanced deepfake dataset towards fair deepfake detection. *arXiv* **2022**, arXiv:2207.10246.
149. GBDF Dataset Download Link. Available online: https://github.com/aakash4305/~GBDF/releases/tag/v1.0 (accessed on 18 April 2024).
150. Cai, Z.; Stefanov, K.; Dhall, A.; Hayat, M. Do you really mean that? Content driven audio-visual deepfake dataset and multimodal method for temporal forgery localization. In Proceedings of the 2022 International Conference on Digital Image Computing: Techniques and Applications (DICTA), Sydney, Australia, 30 November 2022; pp. 1–10.
151. LAVDF Dataset Download Link. Available online: https://drive.google.com/file/d/1-OQ-NDtdEyqHNLaZU1Lt9Upk5wVqfYJw/view (accessed on 25 April 2024).
152. Beniaguev, D. Synthetic Faces High Quality (SFHQ) Dataset. 2022. Available online: https://github.com/SelfishGene/SFHQ-dataset (accessed on 1 January 2024).
153. Boato, G.; Pasquini, C.; Stefani, A.L.; Verde, S.; Miorandi, D. TrueFace: A dataset for the detection of synthetic face images from social networks. In Proceedings of the 2022 IEEE International Joint Conference on Biometrics (IJCB), Abu Dhabi, UAE, 10 October 2022; pp. 1–7.
154. TrueFace Dataset Download Link. Available online: https://drive.google.com/file/d/1WgBrmuKUaLM3YT_5bSgyYUgIUYI_ghOo/view (accessed on 18 April 2024).
155. Park, G.W.; Park, E.J.; Woo, S.S. Zoom-DF: A dataset for video conferencing deepfake. In Proceedings of the 1st Workshop on Security Implications of Deepfakes and Cheapfakes, Nagasaki, Japan, 30 May 2022; pp. 7–11.
156. AV-Deepfake1M Dataset Download Link. Available online: https://monashuni-my.sharepoint.com/personal/zhixi_cai_monash_edu/_layouts/15/onedrive.aspx?id=%2Fpersonal%2Fzhixi%5Fcai%5Fmonash%5Fedu%2FDocuments%2FDataset%2FAV%2DDeepfake1M&ga=1 (accessed on 18 April 2024).
157. Wang, S.; Zhu, Y.; Wang, R.; Dharmasiri, A.; Russakovsky, O.; Wu, Y. DETER: Detecting Edited Regions for Deterring Generative Manipulations. *arXiv* **2023**, arXiv:2312.10539.
158. DETER Dataset Download Link. Available online: https://deter2024.github.io/deter/ (accessed on 18 April 2024).
159. Alnaim, N.M.; Almutairi, Z.M.; Alsuwat, M.S.; Alalawi, H.H.; Alshobaili, A.; Alenezi, F.S. DFFMD: A Deepfake Face Mask Dataset for Infectious Disease Era with Deepfake Detection Algorithms. *IEEE Access* **2023**, *11*, 16711–16722. [CrossRef]

160. DFFMD Dataset Download Link. Available online: https://www.kaggle.com/datasets/hhalalwi/deepfake-face-mask-dataset-dffmd (accessed on 18 April 2024).
161. Narayan, K.; Agarwal, H.; Thakral, K.; Mittal, S.; Vatsa, M.; Singh, R. DF-Platter: Multi-Face Heterogeneous Deepfake Dataset. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 9739–9748.
162. Felouat, H.; Nguyen, H.H.; Le, T.N.; Yamagishi, J.; Echizen, I. eKYC-DF: A Large-Scale Deepfake Dataset for Developing and Evaluating eKYC Systems. *IEEE Access* **2024**, *12*, 30876–30892. [CrossRef]
163. Xu, J.; Chen, J.; Song, X.; Han, F.; Shan, H.; Jiang, Y. Identity-Driven Multimedia Forgery Detection via Reference Assistance. *arXiv* **2024**, arXiv:2401.11764.
164. Hou, Y.; Fu, H.; Chen, C.; Li, Z.; Zhang, H.; Zhao, J. PolyGlotFake: A Novel Multilingual and Multimodal DeepFake Dataset. *arXiv* **2024**, arXiv:2405.08838.
165. Ying, Q.; Liu, J.; Li, S.; Xu, H.; Qian, Z.; Zhang, X. RetouchingFFHQ: A Large-scale Dataset for Fine-grained Face Retouching Detection. In Proceedings of the 31st ACM International Conference on Multimedia, Ottawa, ON, Canada, 29 October 2023; pp. 737–746.
166. Retouching FFHQ Dataset Download Link. Available online: https://drive.google.com/drive/folders/194Viqm8Xh8qleYf66kdSIcGVRupUOYvN (accessed on 18 April 2024).
167. Cho, B.; Le, B.M.; Kim, J.; Woo, S.; Tariq, S.; Abuadbba, A.; Moore, K. Towards Understanding of Deepfake Videos in the Wild. In Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, Birmingham UK, 21 October 2023; pp. 4530–4537.
168. Nowroozi, E.; Habibi, Y.; Conti, M. Spritz-PS: Validation of Synthetic Face Images Using a Large Dataset of Printed Documents. *arXiv* **2023**, arXiv:2304.02982.
169. Spritz-PS Dataset Download Link. Available online: https://ieee-dataport.org/documents/spritz-ps-validation-synthetic-face-images-using-large-dataset-printed-documents (accessed on 24 April 2024).
170. Bei, Y.; Lou, H.; Geng, J.; Liu, E.; Cheng, L.; Song, J.; Song, M.; Feng, Z. A Large-scale Universal Evaluation Benchmark For Face Forgery Detection. *arXiv* **2024**, arXiv:2406.09181.
171. DeepFaceGen Link. 2024. Available online: https://github.com/HengruiLou/DeepFaceGen (accessed on 20 June 2024).
172. Yan, Z.; Yao, T.; Chen, S.; Zhao, Y.; Fu, X.; Zhu, J.; Luo, D.; Yuan, L.; Wang, C.; Ding, S.; et al. DF40: Toward Next-Generation Deepfake Detection. *arXiv* **2024**, arXiv:2406.13495.
173. Abu-El-Haija, S.; Kothari, N.; Lee, J.; Natsev, P.; Toderici, G.; Varadarajan, B.; Vijayanarasimhan, S. Youtube-8m: A large-scale video classification benchmark. *arXiv* **2016**, arXiv:1609.08675.
174. Deepfakes Code Link. Available online: https://github.com/deepfakes/faceswap (accessed on 20 June 2024).
175. Fakeapp Link. Available online: https://www.fakeapp.com/ (accessed on 7 January 2023).
176. Thies, J.; Zollhöfer, M.; Nießner, M. Deferred neural rendering: Image synthesis using neural textures. *ACM Trans. Graph. (TOG)* **2019**, *38*, 1–2. [CrossRef]
177. DeepFaceLab. Available online: https://github.com/iperov/DeepFaceLab, (accessed on 8 January 2024).
178. Choi, Y.; Choi, M.; Kim, M.; Ha, J.W.; Kim, S.; Choo, J. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8789–8797.
179. Zakharov, E.; Shysheya, A.; Burkov, E.; Lempitsky, V. Few-shot adversarial learning of realistic neural talking head models. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9459–9468.
180. Nirkin, Y.; Keller, Y.; Hassner, T. Fsgan: Subject agnostic face swapping and reenactment. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 7184–7193.
181. Polyak, A.; Wolf, L.; Taigman, Y. TTS skins: Speaker conversion via ASR. *arXiv* **2019**, arXiv:1904.08983.
182. Perov, I.; Gao, D.; Chervoniy, N.; Liu, K.; Marangonda, S.; Umé, C.; Dpfks, M.; Facenheim, C.S.; Luis, R.P.; Jiang, J.; et al. DeepFaceLab: Integrated, flexible and extensible face-swapping framework. *arXiv* **2020**, arXiv:2005.05535.
183. Faceswap. Available online: https://github.com/MarekKowalski/FaceSwap/ (accessed on 7 January 2024).
184. Wang, T.C.; Liu, M.Y.; Zhu, J.Y.; Liu, G.; Tao, A.; Kautz, J.; Catanzaro, B. Video-to-video synthesis. *arXiv* **2018**, arXiv:1808.06601.
185. Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B. The cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3213–3223.
186. Xu, Z.; Yu, X.; Hong, Z.; Zhu, Z.; Han, J.; Liu, J.; Ding, E.; Bai, X. Facecontroller: Controllable attribute editing for face in the wild. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 18 May 2021; Volume 35, pp. 3083–3091.
187. Prajwal, K.R.; Mukhopadhyay, R.; Namboodiri, V.P.; Jawahar, C.V. 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, Seattle, WA, USA, 12–16 October 2020; pp. 484–492.
188. Korshunova, I.; Shi, W.; Dambre, J.; Theis, L. Fast face-swap using convolutional neural networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3677–3685.

189. Lee, C.H.; Liu, Z.; Wu, L.; Luo, P. Maskgan: Towards diverse and interactive facial image manipulation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 5549–5558.
190. Choi, Y.; Uh, Y.; Yoo, J.; Ha, J.W. Stargan v2: Diverse image synthesis for multiple domains. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 8188–8197.
191. Chen, L.; Maddox, R.K.; Duan, Z.; Xu, C. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7832–7841.
192. Jo, Y.; Park, J. Sc-fegan: Face editing generative adversarial network with user's sketch and color. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1745–1753.
193. Ye, Z.; Xia, M.; Yi, R.; Zhang, J.; Lai, Y.K.; Huang, X.; Zhang, G.; Liu, Y.J. Audio-driven talking face video generation with dynamic convolution kernels. *IEEE Trans. Multimed.* **2022**, *25*, 2033–2046. [CrossRef]
194. Pidhorskyi, S.; Adjeroh, D.A.; Doretto, G. Adversarial latent autoencoders. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 14104–14113.
195. Shen, Y.; Gu, J.; Tang, X.; Zhou, B. Interpreting the latent space of gans for semantic face editing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 9243–9252.
196. Pérez, P.; Gangnet, M.; Blake, A. Poisson image editing. In *Seminal Graphics Papers: Pushing the Boundaries*; Association for Computing Machinery: New York, NY, USA, 2023; Volume 2, pp. 577–582.
197. Viazovetskyi, Y.; Ivashkin, V.; Kashin, E. Stylegan2 distillation for feed-forward image manipulation. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XXII 16; pp. 170–186.
198. Kim, H.; Garrido, P.; Tewari, A.; Xu, W.; Thies, J.; Niessner, M.; Pérez, P.; Richardt, C.; Zollhöfer, M.; Theobalt, C. Deep video portraits. *ACM Trans. Graph. (TOG)* **2018**, *37*, 1–4. [CrossRef]
199. Brock, A.; Donahue, J.; Simonyan, K. Large scale GAN training for high fidelity natural image synthesis. *arXiv* **2018**, arXiv:1809.11096.
200. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2223–2232.
201. Park, T.; Liu, M.Y.; Wang, T.C.; Zhu, J.Y. Semantic image synthesis with spatially-adaptive normalization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 2337–2346.
202. Kingma, D.P.; Dhariwal, P. Glow: Generative flow with invertible 1x1 convolutions. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 3–8 December 2018; pp. 10236–10245 .
203. Chen, Q.; Koltun, V. Photographic image synthesis with cascaded refinement networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1511–1520.
204. Li, K.; Zhang, T.; Malik, J. Diverse image synthesis from semantic layouts via conditional imle. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 4220–4229.
205. Dai, T.; Cai, J.; Zhang, Y.; Xia, S.T.; Zhang, L. Second-order attention network for single image super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 11065–11074.
206. Yan, W.; Zhang, Y.; Abbeel, P.; Srinivas, A. Videogpt: Video generation using vq-vae and transformers. *arXiv* **2021**, arXiv:2104.10157.
207. Tian, Y.; Ren, J.; Chai, M.; Olszewski, K.; Peng, X.; Metaxas, D.N.; Tulyakov, S. A good image generator is what you need for high-resolution video synthesis. *arXiv* **2021**, arXiv:2104.15069.
208. Yu, S.; Tack, J.; Mo, S.; Kim, H.; Kim, J.; Ha, J.W.; Shin, J. Generating videos with dynamics-aware implicit generative adversarial networks. *arXiv* **2022**, arXiv:2202.10571.
209. Skorokhodov, I.; Tulyakov, S.; Elhoseiny, M. Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 3626–3636.
210. Dfaker. Available online: https://github.com/dfaker/df (accessed on 8 January 2024).
211. Chan, C.; Ginosar, S.; Zhou, T.; Efros, A.A. Everybody dance now. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 5933–5942.
212. Wang, T.C.; Liu, M.Y.; Zhu, J.Y.; Tao, A.; Kautz, J.; Catanzaro, B. High-resolution image synthesis and semantic manipulation with conditional gans. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8798–8807.
213. Bansal, A.; Ma, S.; Ramanan, D.; Sheikh, Y. Recycle-gan: Unsupervised video retargeting. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 119–135.
214. Wang, J.; Qian, X.; Zhang, M.; Tan, R.T.; Li, H. Seeing What You Said: Talking Face Generation Guided by a Lip Reading Expert. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 14653–14662.
215. Kim, J.; Kong, J.; Son, J. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In Proceedings of the International Conference on Machine Learning, Virtual, 1 July 2021; pp. 5530–5540.
216. ChatGPT. Available online: https://chat.openai.com/ (accessed on 8 January 2024).

217. Casanova, E.; Weber, J.; Shulby, C.D.; Junior, A.C.; Gölge, E.; Ponti, M.A. Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone. In Proceedings of the International Conference on Machine Learning, Baltimore, MD, USA, 28 June 2022; pp. 2709–2720.
218. Liu, Z.; Li, M.; Zhang, Y.; Wang, C.; Zhang, Q.; Wang, J.; Nie, Y. Fine-Grained Face Swapping via Regional GAN Inversion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 8578–8587.
219. Zhao, W.; Rao, Y.; Shi, W.; Liu, Z.; Zhou, J.; Lu, J. DiffSwap: High-Fidelity and Controllable Face Swapping via 3D-Aware Masked Diffusion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 8568–8577.
220. Li, W.; Lin, Z.; Zhou, K.; Qi, L.; Wang, Y.; Jia, J. Mat: Mask-aware transformer for large hole image inpainting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 10758–10768.
221. Xia, B.; Zhang, Y.; Wang, S.; Wang, Y.; Wu, X.; Tian, Y.; Yang, W.; Van Gool, L. Diffir: Efficient diffusion model for image restoration. *arXiv* **2023**, arXiv:2303.09472.
222. Chen, R.; Chen, X.; Ni, B.; Ge, Y. Simswap: An efficient framework for high fidelity face swapping. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12 October 2020; pp. 2003–2011.
223. Rosberg, F.; Aksoy, E.E.; Alonso-Fernandez, F.; Englund, C. FaceDancer: Pose-and occlusion-aware high fidelity face swapping. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 2–7 January 2023; pp. 3454–3463.
224. Groshev, A.; Maltseva, A.; Chesakov, D.; Kuznetsov, A.; Dimitrov, D. GHOST—A new face swap approach for image and video domains. *IEEE Access* **2022**, *10*, 83452–83462. [CrossRef]
225. Deep Insight. Insightface. Available online: https://github.com/deepinsight/insightface (accessed on 28 April 2024).
226. s0md3v. Roop. Available online: https://github.com/s0md3v/roop (accessed on 28 April 2024).
227. Gao, G.; Huang, H.; Fu, C.; Li, Z.; He, R. Information bottleneck disentanglement for identity swapping. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 3404–3413.
228. Betker, J. Tortoise Text-to-Speech. Available online: https://github.com/neonbjb/tortoise-tts (accessed on 28 April 2024).
229. RVC-Project. Rvc: Retrieval-Based Voice Conversion Webui. 2023. Available online: https://github.com/RVC-Project/Retrieval-based-Voice-Conversion-WebUI (accessed on 15 June 2024).
230. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
231. AI, S. Github Repository for Suno Ai's Bark Project. 2023. Available online: https://github.com/suno-ai/bark (accessed on 14 June 2024).
232. Li, J.; Tu, W.; Xiao, L. Freevc: Towards high-quality text-free one-shot voice conversion. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; pp. 1–5.
233. Microsoft: Microsoft Azure Text-To-Speech Services. 2023. Available online: https://azure.microsoft.com/en-us/products/ai-services/text-to-speech (accessed on 14 June 2024).
234. AI, C. Github Repository for Coqui Ai Text-To-Speech. 2023. Available online: https://github.com/coqui-ai/tts (accessed on 14 June 2024).
235. Wang, Y.; Skerry-Ryan, R.J.; Stanton, D.; Wu, Y.; Weiss, R.J.; Jaitly, N.; Yang, Z.; Xiao, Y.; Chen, Z.; Bengio, S.; et al. Tacotron: Towards end-to-end speech synthesis. *arXiv* **2017**, arXiv:1703.10135.
236. Wang, C.; Chen, S.; Wu, Y.; Zhang, Z.; Zhou, L.; Liu, S.; Chen, Z.; Liu, Y.; Wang, H.; Li, J.; et al. Neural codec language models are zero-shot text to speech synthesizers. *arXiv* **2023**, arXiv:2301.02111.
237. Cheng, K.; Cun, X.; Zhang, Y.; Xia, M.; Yin, F.; Zhu, M.; Wang, X.; Wang, J.; Wang, N. Videoretalking: Audio-based lip synchronization for talking head video editing in the wild. In Proceedings of the SIGGRAPH Asia 2022 Conference Papers, Daegu, Republic of Korea, 6–9 December 2022; pp. 1–9.
238. Megvii Face++, Megvii Face Retouching API. Available online: https://www.faceplusplus.com.cn/sdk/facebeautify/ (accessed on 8 January 2024).
239. Alibaba Cloud, Alibaba Face Retouching API. Available online: https://help.aliyun.com/document_detail/159210.html (accessed on 8 January 2024).
240. Tencent Cloud, Tencent Face Retouching API. Available online: https://cloud.tencent.com/document/product/1172/40715 (accessed on 8 January 2024).
241. DeepFaceLive Link 1. Available online: https://drive.google.com/file/d/1KS37b2IBuljJuZiJsgnWuzs7Y5OfkOyI/view/ (accessed on 8 January 2024).
242. DeepFaceLive Link 2. Available online: https://github.com/iperov/DeepFaceLive (accessed on 8 January 2024).
243. FacePlay App. Available online: https://www.faceplay.cc/ (accessed on 8 January 2024).
244. Reface App Link 1. Available online: https://apps.apple.com/us/app/reface-face-swap-ai-photo-app/id1488782587 (accessed on 8 January 2024).
245. Reface App Link 2. Available online: https://reface.app/ (accessed on 8 January 2024).
246. Reface App Link 3. Available online: https://play.google.com/store/apps/details?id=video.reface.app&hl=en_US&gl=US (accessed on 8 January 2024).

247. DeepfakeStudio App. Available online: https://play.google.com/store/apps/details?id=com.deepworkings.dfstudio&hl=en&gl=US&pli=1/ (accessed on 8 January 2024).
248. Revive. Available online: https://play.google.com/store/apps/details?id=revive.app&hl=en_US&gl=US (accessed on 8 January 2024).
249. LicoLico App. Available online: http://licolico.cn/home/ (accessed on 8 January 2024).
250. Fakeit App. Available online: https://vk.com/fakeit/ (accessed on 8 January 2024).
251. DeepFaker App. Available online: https://deepfaker.app/ (accessed on 8 January 2024).
252. DeepFakesWeb Site. Available online: https://deepfakesweb.com/ (accessed on 8 January 2024).
253. Deepcake.io Link 1. Available online: http://deepcake.io/ (accessed on 8 January 2024).
254. Deepcake.io Link 2. Available online: https://www.instagram.com/deepcake.io/ (accessed on 8 January 2024).
255. DeepFaker Bot Site. Available online: https://t.me/DeepFakerBot/ (accessed on 8 January 2024).
256. Revel.ai Site. Available online: http://revel.ai/ (accessed on 8 January 2024).
257. Shiohara, K.; Yang, X.; Taketomi, T. BlendFace: Re-designing Identity Encoders for Face-Swapping. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–3 October 2023; pp. 7634–7644.
258. Fried, O.; Tewari, A.; Zollhöfer, M.; Finkelstein, A.; Shechtman, E.; Goldman, D.B.; Genova, K.; Jin, Z.; Theobalt, C.; Agrawala, M. Text-based editing of talking-head video. *ACM Trans. Graph. (TOG)* **2019**, *38*, 1–14. [CrossRef]
259. Siarohin, A.; Roy, S.; Lathuilière, S.; Tulyakov, S.; Ricci, E.; Sebe, N. Motion-supervised co-part segmentation. In Proceedings of the IEEE International Conference on Pattern Recognition (ICPR), Virtual, 10–15 January 2021; pp. 9650–9657.
260. Deng, Y.; Yang, J.; Chen, D.; Wen, F.; Tong, X. Disentangled and controllable face image generation via 3d imitative-contrastive learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 14–19 June 2020; pp. 5154–5163.
261. PromptHero. Openjourney. 2023. Available online: http://openjourney.art/ (accessed on 20 June 2024).
262. Stability.ai, Stable Diffusion. 2023. Available online: https://stability.ai/ (accessed on 20 June 2024).
263. Baidu, Wenxin. 2022. Available online: https://yige.baidu.com/ (accessed on 20 June 2024).
264. Midjourney, Midjourney. 2022. Available online: https://www.midjourney.com/home (accessed on 20 June 2024).
265. Tao, M.; Tang, H.; Wu, F.; Jing, X.Y.; Bao, B.K.; Xu, C. Df-gan: A simple and effective baseline for text-to-image synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 16515–16525.
266. Open AI, DALL·E. 2020. Available online: https://openai.com/index/dall-e-3 (accessed on 20 June 2024).
267. Lin, S.; Yang, X. Animatediff-lightning: Cross-model diffusion distillation. *arXiv* **2024**, arXiv:2403.12706.
268. Wang, F.Y.; Huang, Z.; Shi, X.; Bian, W.; Song, G.; Liu, Y.; Li, H. AnimateLCM: Accelerating the Animation of Personalized Diffusion Models and Adapters with Decoupled Consistency Learning. *arXiv* **2024**, arXiv:2402.00769.
269. Mullan, J.; Crawbuck, D.; Sastry, A. Hotshot-XL. 2023. Available online: https://github.com/hotshotco/hotshot-xl (accessed on 20 June 2024).
270. Academy for Discovery, Adventure, Momentum and Outlook. Zeroscope. 2023. Available online: https://huggingface.co/cerspense/zeroscope_v2_576w (accessed on 20 June 2024).
271. Yuan, S.; Huang, J.; Shi, Y.; Xu, Y.; Zhu, R.; Lin, B.; Cheng, X.; Yuan, L.; Luo, J. MagicTime: Time-lapse Video Generation Models as Metamorphic Simulators. *arXiv* **2024**, arXiv:2404.05014.
272. Wang, H. Inswapper. Available online: https://github.com/haofanwang/inswapper (accessed on 3 July 2024).
273. Xu, C.; Zhang, J.; Han, Y.; Tian, G.; Zeng, X.; Tai, Y.; Wang, Y.; Wang, C.; Liu, Y. Designing one unified framework for high-fidelity face reenactment and swapping. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; pp. 54–71.
274. Wang, T.C.; Liu, M.Y.; Tao, A.; Liu, G.; Kautz, J.; Catanzaro, B. Few-shot video-to-video synthesis. *arXiv* **2019**, arXiv:1910.12713.
275. Siarohin, A.; Woodford, O.J.; Ren, J.; Chai, M.; Tulyakov, S. Motion representations for articulated animation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 13653–13662.
276. Wang, T.C.; Mallya, A.; Liu, M.Y. One-shot free-view neural talking-head synthesis for video conferencing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 10039–10049.
277. Ren, Y.; Li, G.; Chen, Y.; Li, T.H.; Liu, S. Pirenderer: Controllable portrait image generation via semantic neural rendering. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 14–19 June 2021; pp. 13759–13768.
278. Zhao, J.; Zhang, H. Thin-plate spline motion model for image animation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 3657–3666.
279. Wang, Y.; Yang, D.; Bremond, F.; Dantcheva, A. Latent image animator: Learning to animate images via latent space navigation. *arXiv* **2022**, arXiv:2203.09043.
280. Hong, F.T.; Zhang, L.; Shen, L.; Xu, D. Depth-aware generative adversarial network for talking head video generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 3397–3406.
281. Zhang, W.; Cun, X.; Wang, X.; Zhang, Y.; Shen, X.; Guo, Y.; Shan, Y.; Wang, F. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 8652–8661.

282. Hong, F.T.; Xu, D. Implicit identity representation conditioned memory compensation network for talking head video generation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 23062–23072.
283. Bounareli, S.; Tzelepis, C.; Argyriou, V.; Patras, I.; Tzimiropoulos, G. Hyperreenact: One-shot reenactment via jointly learning to refine and retarget faces. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 7149–7159.
284. HeyGen. Available online: https://www.heygen.com (accessed on 3 July 2024).
285. Esser, P.; Rombach, R.; Ommer, B. Taming transformers for high-resolution image synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 12873–12883.
286. Karras, T.; Aittala, M.; Laine, S.; Härkönen, E.; Hellsten, J.; Lehtinen, J.; Aila, T. Alias-free generative adversarial networks. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 852–863.
287. Sauer, A.; Schwarz, K.; Geiger, A. Stylegan-xl: Scaling stylegan to large diverse datasets. In Proceedings of the ACM SIGGRAPH 2022 Conference Proceedings, Vancouver, BC, Canada, 7–11 August 2022; pp. 1–10.
288. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 10684–10695.
289. Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 6840–6851.
290. Liu, J.; Wang, Q.; Fan, H.; Wang, Y.; Tang, Y.; Qu, L. Residual denoising diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–21 June 2024; pp. 2773–2783.
291. Chen, J.; Yu, J.; Ge, C.; Yao, L.; Xie, E.; Wu, Y.; Wang, Z.; Kwok, J.; Luo, P.; Lu, H.; et al. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv* **2023**, arXiv:2310.00426.
292. Peebles, W.; Xie, S. Scalable diffusion models with transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–3 October 2023; pp. 4195–4205.
293. Atito, S.; Awais, M.; Kittler, J. Sit: Self-supervised vision transformer. *arXiv* **2021**, arXiv:2104.03602.
294. Huang, Z.; Chan, K.C.; Jiang, Y.; Liu, Z. Collaborative diffusion for multi-modal face generation and editing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 6080–6090.
295. Tov, O.; Alaluf, Y.; Nitzan, Y.; Patashnik, O.; Cohen-Or, D. Designing an encoder for stylegan image manipulation. *ACM Trans. Graph.* **2021**, *40*, 1–14. [CrossRef]
296. Xie, L.; Wang, X.; Zhang, H.; Dong, C.; Shan, Y. Vfhq: A high-quality dataset and benchmark for video face super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 657–666.
297. Liu, Z.; Luo, P.; Wang, X.; Tang, X. Large-Scale CelebFaces Attributes (CelebA) Dataset. Available online: https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html (accessed on 15 August 2018).
298. Li, L.; Bao, J.; Yang, H.; Chen, D.; Wen, F. Advancing high fidelity identity swapping for forgery detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 14–19 June 2020; pp. 5074–5083.
299. Arik, S.; Chen, J.; Peng, K.; Ping, W.; Zhou, Y. Neural voice cloning with a few samples. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 1–11.
300. Baidu Dataset Download Link. Available online: https://audiodemos.github.io/ (accessed on 11 April 2024).
301. Solak, I. The M-AILABS Speech Dataset. Available online: https://www.caito.de/2019/01/the-m-ailabs-speech-dataset/ (accessed on 4 January 2024).
302. Kondratiuk, S.; Hartvih, D.; Krak, I.; Barmak, O.; Kuznetsov, V. Deep Learning Models for Ukrainian Text to Speech Synthesis. In Proceedings of the 4th International Workshop on Intelligent Information Technologies and Systems of Information Security, Khmelnytskyi, Ukraine, 22–24 March 2023; pp. 1–11. Available online: https://ceur-ws.org/Vol-3373/paper10.pdf (accessed on 4 July 2024).
303. Yamagishi, J.; Todisco, M.; Sahidullah, M.; Delgado, H.; Wang, X.; Evans, N.; Kinnunen, T.; Lee, K.A.; Vestman, V.; Nautsch, A. Asvspoof 2019: Automatic Speaker Verification Spoofing and Countermeasures Challenge Evaluation Plan. Available online: http://www.asvspoof.org/asvspoof2019/asvspoof2019evaluationplan.pdf (accessed on 15 January 2019).
304. Asvspoof 2019 Download Link. Available online: https://datashare.ed.ac.uk/handle/10283/3336 (accessed on 17 April 2024).
305. Lieto, A.; Moro, D.; Devoti, F.; Parera, C.; Lipari, V.; Bestagini, P.; Tubaro, S. Hello? Who Am I Talking to? A Shallow CNN Approach for Human vs. Bot Speech Classification. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12 May 2019; pp. 2577–2581.
306. Mari, D.; Salvi, D.; Bestagini, P.; Milani, S. All-for-One and One-For-All: Deep learning-based feature fusion for Synthetic Speech Detection. *arXiv* **2023**, arXiv:2307.15555.
307. Borrelli, C. Data Driven and Signal Processing Techniques for Audio Forensics. 2022, pp. 1–190. Available online: https://www.politesi.polimi.it/handle/10589/188972 (accessed on 28 April 2024).
308. Reimao, R.; Tzerpos, V. FoR: A dataset for synthetic speech detection. In Proceedings of the 2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD), Timisoara, Romania, 10 October 2019; pp. 1–10.
309. Fake or Real (FoR) Dataset Link. Available online: https://bil.eecs.yorku.ca/datasets/ (accessed on 17 April 2024).
310. Ballesteros, D.M.; Rodriguez, Y.; Renza, D. A dataset of histograms of original and fake voice recordings (H-Voice). *Data Brief* **2020**, *29*, 105331. [CrossRef]

311. H-Voice Dataset Link 1. Available online: https://www.kaggle.com/datasets/dduongtrandai/hvoice-fake-voice (accessed on 17 April 2024).
312. H-Voice Dataset Link 2. Available online: https://data.mendeley.com/datasets/ytkv9w92t6/1 (accessed on 17 April 2024).
313. H-Voice Dataset Link 3. Available online: https://data.mendeley.com/datasets/k47yd3m28w/4 (accessed on 17 April 2024).
314. Lataifeh, M.; Elnagar, A. Ar-DAD: Arabic diversified audio dataset. *Data Brief* **2020**, *33*, 106503. [CrossRef]
315. ARDAD2020. Available online: https://data.mendeley.com/datasets/3kndp5vs6b/1 (accessed on 17 April 2024).
316. Zhao, Y.; Huang, W.C.; Tian, X.; Yamagishi, J.; Das, R.K.; Kinnunen, T.; Ling, Z.; Toda, T. Voice conversion challenge 2020: Intra-lingual semi-parallel and cross-lingual voice conversion. *arXiv* **2020**, arXiv:2008.12527.
317. Wang, C.; Yi, J.; Tao, J.; Zhang, C.; Zhang, S.; Chen, X. Detection of Cross-Dataset Fake Audio Based on Prosodic and Pronunciation Features. *arXiv* **2023**, arXiv:2305.13700.
318. VCC Dataset. Available online: https://zenodo.org/records/4345689 (accessed on 17 April 2024).
319. Liu, X.; Wang, X.; Sahidullah, M.; Patino, J.; Delgado, H.; Kinnunen, T.; Todisco, M.; Yamagishi, J.; Evans, N.; Nautsch, A.; et al. Asvspoof 2021: Towards spoofed and deepfake speech detection in the wild. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2023**, *31*, 2507–2522. [CrossRef]
320. Yamagishi, J.; Wang, X.; Todisco, M.; Sahidullah, M.; Patino, J.; Nautsch, A.; Liu, X.; Lee, K.A.; Kinnunen, T.; Evans, N.; et al. ASVspoof 2021: Accelerating progress in spoofed and deepfake speech detection. *arXiv* **2021**, arXiv:2109.00537.
321. Asvspoof2021 Link 1. Available online: https://zenodo.org/records/4835108 (accessed on 17 April 2024).
322. Asvspoof2021 Link 2. Available online: https://zenodo.org/records/4834716 (accessed on 17 April 2024).
323. Asvspoof2021 Link 3. Available online: https://zenodo.org/records/4837263 (accessed on 17 April 2024).
324. Zhang, Z.; Gu, Y.; Yi, X.; Zhao, X. FMFCC-A: A challenging Mandarin dataset for synthetic speech detection. In *International Workshop on Digital Watermarking*; Springer International Publishing: Cham, Switzerland, 2021; pp. 117–131.
325. FMFCC-A Dataset Link. Available online: https://pan.baidu.com/s/1CGPkC8VfjXVBZjluEHsW6g (accessed on 17 April 2024).
326. Frank, J.; Schönherr, L. Wavefake: A data set to facilitate audio deepfake detection. *arXiv* **2021**, arXiv:2111.02813.
327. WaveFake Dataset Link. Available online: https://zenodo.org/records/5642694 (accessed on 17 April 2024).
328. Yi, J.; Fu, R.; Tao, J.; Nie, S.; Ma, H.; Wang, C.; Wang, T.; Tian, Z.; Bai, Y.; Fan, C.; et al. Add 2022: The first audio deep synthesis detection challenge. In Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 22–27 May 2022; pp. 9216–9220.
329. ADD 2022 Dataset Download Link. Available online: http://addchallenge.cn/download (accessed on 25 April 2024).
330. Ma, H.; Yi, J.; Wang, C.; Yan, X.; Tao, J.; Wang, T.; Wang, S.; Xu, L.; Fu, R. FAD: A Chinese dataset for fake audio detection. *arXiv* **2022**, arXiv:2207.12308.
331. Chinese Fake Audio Dataset Download Link. Available online: https://zenodo.org/records/6635521 (accessed on 25 April 2024).
332. Müller, N.M.; Czempin, P.; Dieckmann, F.; Froghyar, A.; Böttinger, K. Does audio deepfake detection generalize? *arXiv* **2022**, arXiv:2203.16263.
333. In-The-Wild Dataset Download Link. Available online: https://owncloud.fraunhofer.de/index.php/s/JZgXh0JEAF0elxa (accessed on 25 April 2024).
334. Papastergiopoulos, C.; Vafeiadis, A.; Papadimitriou, I.; Votis, K.; Tzovaras, D. On the generalizability of two-dimensional convolutional neural networks for fake speech detection. In Proceedings of the 1st International Workshop on Multimedia AI against Disinformation, Newark, NJ, USA, 27–30 June 2022; pp. 3–9.
335. TIMIT Dataset Download Link. Available online: https://conradsanderson.id.au/vidtimit/#downloads (accessed on 25 April 2024).
336. Yi, J.; Tao, J.; Fu, R.; Yan, X.; Wang, C.; Wang, T.; Zhang, C.Y.; Zhang, X.; Zhao, Y.; Ren, Y.; et al. ADD 2023: The Second Audio Deepfake Detection Challenge. *arXiv* **2023**, arXiv:2305.13774.
337. Xie, Y.; Zhou, J.; Lu, X.; Jiang, Z.; Yang, Y.; Cheng, H.; Ye, L. FSD: An initial chinese dataset for fake song detection. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Republic of Korea, 14–19 April 2024; pp. 4605–4609.
338. Fake Song Detection (FSD). Available online: https://github.com/xieyuankun/FSD-Dataset (accessed on 15 June 2024).
339. Half-Truth Dataset Download Link. Available online: https://zenodo.org/records/10377492 (accessed on 25 April 2024).
340. Yan, X.; Yi, J.; Tao, J.; Wang, C.; Ma, H.; Tian, Z.; Fu, R. System fingerprints detection for deepfake audio: An initial dataset and investigation. *arXiv* **2022**, arXiv:2208.10489.
341. Salvi, D.; Hosler, B.; Bestagini, P.; Stamm, M.C.; Tubaro, S. TIMIT-TTS: A Text-to-Speech Dataset for Multimodal Synthetic Media Detection. *IEEE Access* **2023**, *11*, 50851–50866. [CrossRef]
342. TIMIT-TTS Dataset Download Link. Available online: https://zenodo.org/records/6560159 (accessed on 25 April 2024).
343. Li, Y.; Zhang, M.; Ren, M.; Ma, M.; Wei, D.; Yang, H. Cross-Domain Audio Deepfake Detection: Dataset and Analysis. *arXiv* **2024**, arXiv:2404.04904.
344. Xie, Y.; Lu, Y.; Fu, R.; Wen, Z.; Wang, Z.; Tao, J.; Qi, X.; Wang, X.; Liu, Y.; Cheng, H.; et al. The Codecfake Dataset and Countermeasures for the Universally Detection of Deepfake Audio. *arXiv* **2024**, arXiv:2405.04880.
345. Codecfake Dataset. Available online: https://github.com/xieyuankun/Codecfake (accessed on 16 June 2024).
346. Zang, Y.; Shi, J.; Zhang, Y.; Yamamoto, R.; Han, J.; Tang, Y.; Xu, S.; Zhao, W.; Guo, J.; Toda, T.; et al. CtrSVDD: A Benchmark Dataset and Baseline Analysis for Controlled Singing Voice Deepfake Detection. *arXiv* **2024**, arXiv:2406.02438.

347. Xie, Z.; Li, B.; Xu, X.; Liang, Z.; Yu, K.; Wu, M. FakeSound: Deepfake General Audio Detection. *arXiv* **2024**, arXiv:2406.08052.
348. Kim, C.D.; Kim, B.; Lee, H.; Kim, G. Audiocaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*; Association for Computational Linguistics: Minneapolis, MN, USA, 2019; Volume 1, pp. 119–132.
349. FakeSound Samples. Available online: https://fakesounddata.github.io/ (accessed on 20 June 2024).
350. FakeSounds Download Link. Available online: https://drive.google.com/file/d/1Zma4npCeTmdvoNc3xRvQrzpHRlJkRWK4/view (accessed on 20 June 2024).
351. Yi, J.; Wang, C.; Tao, J.; Zhang, C.Y.; Fan, C.; Tian, Z.; Ma, H.; Fu, R. Scenefake: An initial dataset and benchmarks for scene fake audio detection. *Pattern Recognit.* **2024**, *152*, 110468. [CrossRef]
352. Zang, Y.; Zhang, Y.; Heydari, M.; Duan, Z. SingFake: Singing Voice Deepfake Detection. *arXiv* **2023**, arXiv:2309.07525.
353. Wang, X.; Yamagishi, J.; Todisco, M.; Delgado, H.; Nautsch, A.; Evans, N.; Sahidullah, M.; Vestman, V.; Kinnunen, T.; Lee, K.A.; et al. ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech. *Comput. Speech Lang.* **2020**, *64*, 101114. [CrossRef]
354. Todisco, M.; Wang, X.; Vestman, V.; Sahidullah, M.; Delgado, H.; Nautsch, A.; Yamagishi, J.; Evans, N.; Kinnunen, T.; Lee, K.A. ASVspoof 2019: Future horizons in spoofed and fake audio detection. *arXiv* **2019**, arXiv:1904.05441.
355. Yamagishi, J.; Veaux, C.; MacDonald, K. *CSTR VCTK Corpus: English Multi-Speaker Corpus for CSTR Voice Cloning Toolkit (Version 0.92)*; University of Edinburgh; The Centre for Speech Technology Research (CSTR): Edinburgh, UK, 2019.
356. Wu, Z.; Watts, O.; King, S. Merlin: An Open Source Neural Network Speech Synthesis System. In Proceedings of the 9th ISCA Speech Synthesis Workshop, Sunnyvale, CA, USA 13–15 September 2016; pp. 202–207. Available online: https://github.com/CSTR-Edinburgh/merlin (accessed on 28 April 2024).
357. CURRENNT Tool. Available online: https://github.com/nii-yamagishilab/project-CURRENNT-public (accessed on 28 April 2024).
358. Schroder, M.; Charfuelan, M.; Pammi, S.; Steiner, I. Open Source Voice Creation Toolkit for the MARY TTS Platform. Interspeech. 2011, pp. 3253–3256. Available online: https://github.com/marytts/marytts (accessed on 28 April 2024).
359. Kawakami, K. Supervised Sequence Labelling with Recurrent Neural Networks. Doctoral Dissertation, Technical University of Munich, Munich, Germany, 2008. Available online: https://mediatum.ub.tum.de/doc/1289309/document.pdf (accessed on 4 July 2024).
360. Agiomyrgiannakis, Y. Vocaine the vocoder and applications in speech synthesis. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brisbane, Australia, 19–24 April 2015; pp. 4230–4234.
361. Kalchbrenner, N.; Elsen, E.; Simonyan, K.; Noury, S.; Casagrande, N.; Lockhart, E.; Stimberg, F.; Oord, A.; Dieleman, S.; Kavukcuoglu, K. Efficient neural audio synthesis. In Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 2410–2419.
362. Oord, A.V.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A.; Kavukcuoglu, K. Wavenet: A generative model for raw audio. *arXiv* **2016**, arXiv:1609.03499.
363. Tanaka, K.; Kameoka, H.; Kaneko, T.; Hojo, N. WaveCycleGAN2: Time-domain neural post-filter for speech waveform generation. *arXiv* **2019**, arXiv:1904.02892.
364. Amazon AWS Polly. Available online: https://aws.amazon.com/polly/ (accessed on 9 January 2024).
365. Google Cloud Text-to-Speech with Wavenet. Available online: https://cloud.google.com/text-to-speech/ (accessed on 9 January 2024).
366. Microsoft Azure. Available online: https://azure.microsoft.com/en-us/services/cognitive-services/text-to-speech/ (accessed on 9 January 2024).
367. IBM Watson. Available online: www.ibm.com/watson/services/text-to-speech/ (accessed on 9 January 2024).
368. Ping, W.; Peng, K.; Gibiansky, A.; Arik, S.O.; Kannan, A.; Narang, S.; Raiman, J.; Miller, J. Deep voice 3: Scaling text-to-speech with convolutional sequence learning. *arXiv* **2017**, arXiv:1710.07654.
369. Ren, Y.; Hu, C.; Tan, X.; Qin, T.; Zhao, S.; Zhao, Z.; Liu, T.Y. Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv* **2020**, arXiv:2006.04558.
370. Chou, J.C.; Yeh, C.C.; Lee, H.Y. One-shot voice conversion by separating speaker and content representations with instance normalization. *arXiv* **2019**, arXiv:1904.05742.
371. Bińkowski, M.; Donahue, J.; Dieleman, S.; Clark, A.; Elsen, E.; Casagrande, N.; Cobo, L.C.; Simonyan, K. High fidelity speech synthesis with adversarial networks. *arXiv* **2019**, arXiv:1909.11646.
372. Sonobe, R.; Takamichi, S.; Saruwatari, H. JSUT corpus: Free large-scale Japanese speech corpus for end-to-end speech synthesis. *arXiv* **2017**, arXiv:1711.00354.
373. Kumar, K.; Kumar, R.; De Boissiere, T.; Gestin, L.; Teoh, W.Z.; Sotelo, J.; De Brebisson, A.; Bengio, Y.; Courville, A.C. Melgan: Generative adversarial networks for conditional waveform synthesis. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 1–12.
374. Yamamoto, R.; Song, E.; Kim, J.M. Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Virtual, 4–8 May 2020; pp. 6199–6203.
375. Yang, G.; Yang, S.; Liu, K.; Fang, P.; Chen, W.; Xie, L. Multi-band melgan: Faster waveform generation for high-quality text-to-speech. In Proceedings of the 2021 IEEE Spoken Language Technology Workshop (SLT), Virtual, 19 January 2021; pp. 492–498.

376. Kong, J.; Kim, J.; Bae, J. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 17022–17033.
377. Prenger, R.; Valle, R.; Catanzaro, B. Waveglow: A flow-based generative network for speech synthesis. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12 May 2019; pp. 3617–3621.
378. Kawahara, H. STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds. *Acoust. Sci. Technol.* **2006**, *27*, 349–353. [CrossRef]
379. Perraudin, N.; Balazs, P.; Søndergaard, P.L. A fast Griffin-Lim algorithm. In Proceedings of the 2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY, USA, 20 October 2013; pp. 1–4.
380. Valin, J.M.; Skoglund, J. LPCNet: Improving neural speech synthesis through linear prediction. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12 May 2019; pp. 5891–5895.
381. Mustafa, A.; Pia, N.; Fuchs, G. Stylemelgan: An efficient high-fidelity adversarial vocoder with temporal adaptive normalization. In Proceedings of the ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Virtual, 6–11 June 2021; pp. 6034–6038.
382. Morise, M.; Yokomori, F.; Ozawa, K. WORLD: A vocoder-based high-quality speech synthesis system for real-time applications. *IEICE Trans. Inf. Syst.* **2016**, *99*, 1877–1884. [CrossRef]
383. Wang, Y.; Stanton, D.; Zhang, Y.; Ryan, R.S.; Battenberg, E.; Shor, J.; Xiao, Y.; Jia, Y.; Ren, F.; Saurous, R.A. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In Proceedings of the International Conference on Machine Learning, Vienna, Austria, 10–15 July 2018; pp. 5180–5189.
384. Skerry-Ryan, R.J.; Battenberg, E.; Xiao, Y.; Wang, Y.; Stanton, D.; Shor, J.; Weiss, R.; Clark, R.; Saurous, R.A. Towards end-to-end prosody transfer for expressive speech synthesis with tacotron. In Proceedings of the International Conference on Machine Learning, Vienna, Austria, 10–15 July 2018; pp. 4693–4702.
385. Tak, H.; Patino, J.; Todisco, M.; Nautsch, A.; Evans, N.; Larcher, A. End-to-end anti-spoofing with rawnet2. In Proceedings of the ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Virtual, 6 June 2021; pp. 6369–6373.
386. Tak, H.; Jung, J.W.; Patino, J.; Kamble, M.; Todisco, M.; Evans, N. End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection. *arXiv* **2021**, arXiv:2107.12710.
387. Shen, J.; Pang, R.; Weiss, R.J.; Schuster, M.; Jaitly, N.; Yang, Z.; Chen, Z.; Zhang, Y.; Wang, Y.; Skerrv-Ryan, R.; et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15 April 2018; pp. 4779–4783.
388. Garofolo, J.S. Timit acoustic phonetic continuous speech corpus. *Linguist. Data Consort.* **1993**, *1993*, LDC93S1.
389. Zhou, K.; Sisman, B.; Liu, R.; Li, H. Emotional voice conversion: Theory, databases and ESD. *Speech Commun.* **2022**, *137*, 1–8. [CrossRef]
390. Zhou, K.; Sisman, B.; Zhang, M.; Li, H. Converting anyone's emotion: Towards speaker-independent emotional voice conversion. *arXiv* **2020**, arXiv:2005.07025.
391. Zhou, K.; Sisman, B.; Li, H. Limited data emotional voice conversion leveraging text-to-speech: Two-stage sequence-to-sequence training. *arXiv* **2021**, arXiv:2103.16809.
392. Fu, C.; Liu, C.; Ishi, C.T.; Ishiguro, H. Cycletransgan-evc: A cyclegan-based emotional voice conversion model with transformer. *arXiv* **2021**, arXiv:2111.15159.
393. Zhou, K.; Sisman, B.; Liu, R.; Li, H. Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset. In Proceedings of the ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Virtual, 6–11 June 2021; pp. 920–924.
394. Gao, J.; Chakraborty, D.; Tembine, H.; Olaleye, O. Nonparallel emotional speech conversion. *arXiv* **2018**, arXiv:1811.01174.
395. Rizos, G.; Baird, A.; Elliott, M.; Schuller, B. Stargan for emotional speech conversion: Validated by data augmentation of end-to-end emotion recognition. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Virtual, 4–8 May 2020; pp. 3502–3506.
396. Zhou, K.; Sisman, B.; Li, H. Transforming spectrum and prosody for emotional voice conversion with non-parallel training data. *arXiv* **2020**, arXiv:2002.00198.
397. SO-VITS. Available online: https://github.com/svc-develop-team/so-vits-svc (accessed on 15 June 2024).
398. Ziyin, L.; Hartwig, T.; Ueda, M. Neural networks fail to learn periodic functions and how to fix it. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1583–1594.
399. Liu, J.; Li, C.; Ren, Y.; Chen, F.; Zhao, Z. Diffsinger: Singing voice synthesis via shallow diffusion mechanism. *AAAI Conf. Artif. Intell.* **2022**, *36*, 11020–11028. [CrossRef]
400. Aispeech. Available online: https://cloud.aispeech.com/openSource/technology/tts (accessed on 10 January 2024).
401. Sogou. Available online: https://ai.sogou.com/ (accessed on 10 January 2024).
402. Alibaba Cloud. Available online: https://ai.aliyun.com/nls/tts (accessed on 10 January 2024).
403. Baidu Ai Cloud. Available online: https://ai.baidu.com/tech/speech/tts (accessed on 10 January 2024).
404. Databaker. Available online: https://data-baker.com/specs/compose/online (accessed on 10 January 2024).

405. Tencent Cloud. Available online: https://cloud.tencent.com/product/tts (accessed on 10 January 2024).
406. iFLYTEK. Available online: https://www.xfyun.cn/services/online_tts (accessed on 10 January 2024).
407. Sanderson, C.; Lovell, B.C. Multi-region probabilistic histograms for robust and scalable identity inference. In Proceedings of the Third International Conference on Advances in Biometrics, ICB 2009, Alghero, Italy, 2–5 June 2009; pp. 199–208.
408. Kim, J.; Kim, S.; Kong, J.; Yoon, S. Glow-tts: A generative flow for text-to-speech via monotonic alignment search. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 8067–8077.
409. Lancucki, A. Fastpitch: Parallel text-to-speech with pitch prediction. In Proceedings of the ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Virtual, 6–11 June 2021; pp. 6588–6592.
410. Beliaev, S.; Rebryk, Y.; Ginsburg, B. TalkNet: Fully-convolutional non-autoregressive speech synthesis model. *arXiv* **2020**, arXiv:2005.05514.
411. Tatanov, O.; Beliaev, S.; Ginsburg, B. Mixer-TTS: Non-autoregressive, fast and compact text-to-speech model conditioned on language model embeddings. In Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Virtual, 22–27 May 2022; pp. 7482–7486.
412. Vainer, J.; Dušek, O. Speedyspeech: Efficient neural speech synthesis. *arXiv* **2020**, arXiv:2008.03802.
413. Durette, P.N. GTTS. Available online: https://github.com/pndurette/gTTS (accessed on 10 January 2024).
414. Silero Team. Silero Models: Pre-Trained Enterprise-Grade stt/tts Models and Benchmarks. Available online: https://github.com/snakers4/silero-models (accessed on 10 January 2024).
415. Müller, M.; Özer, Y.; Krause, M.; Prätzlich, T.; Driedger, J. Sync toolbox: A python package for efficient, robust, and accurate music synchronization. *J. Open Source Softw.* **2021**, *6*, 3434. [CrossRef]
416. Kharitonov, E.; Vincent, D.; Borsos, Z.; Marinier, R.; Girgin, S.; Pietquin, O.; Sharifi, M.; Tagliasacchi, M.; Zeghidour, N. Speak, read and prompt: High-fidelity text-to-speech with minimal supervision. *Trans. Assoc. Comput. Linguist.* **2023**, *11*, 1703–1718. [CrossRef]
417. Barrault, L.; Chung, Y.A.; Meglioli, M.C.; Dale, D.; Dong, N.; Duppenthaler, M.; Duquenne, P.A.; Ellis, B.; Elsahar, H.; Haaheim, J.; et al. Seamless: Multilingual Expressive and Streaming Speech Translation. *arXiv* **2023**, arXiv:2312.05187.
418. Qin, Z.; Zhao, W.; Yu, X.; Sun, X. OpenVoice: Versatile Instant Voice Cloning. *arXiv* **2023**, arXiv:2312.01479.
419. Shi, Y.; Bu, H.; Xu, X.; Zhang, S.; Li, M. AISHELL-3: A Multi-Speaker Mandarin TTS Corpus. In Proceedings of the INTERSPEECH 2021, Brno, Czechia, 30 August–3 September 2021; pp. 2756–2760. Available online: https://www.isca-archive.org/interspeech_2021/shi21c_interspeech.pdf (accessed on 10 January 2024).
420. Zeghidour, N.; Luebs, A.; Omran, A.; Skoglund, J.; Tagliasacchi, M. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *30*, 495–507. [CrossRef]
421. Zhang, X.; Zhang, D.; Li, S.; Zhou, Y.; Qiu, X. SpeechTokenizer: Unified Speech Tokenizer for Speech Language Models. In Proceedings of the Twelfth International Conference on Learning Representations, Vienna, Austria, 7–11 May 2024; pp. 1–21.
422. Du, Z.; Zhang, S.; Hu, K.; Zheng, S. Funcodec: A fundamental, reproducible and integrable open-source toolkit for neural speech codec. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Republic of Korea, 14–19 April 2024; pp. 591–595.
423. Defossez, A.; Copet, J.; Synnaeve, G.; Adi, Y. High fidelity neural audio compression. *arXiv* **2022**, arXiv:2210.13438.
424. Wu, Y.C.; Gebru, I.D.; Markovic, D.; Richard, A. Audiodec: An open-source streaming high-fidelity neural audio codec. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; pp. 1–5.
425. Yang, D.; Liu, S.; Huang, R.; Tian, J.; Weng, C.; Zou, Y. Hifi-codec: Group-residual vector quantization for high fidelity audio codec. *arXiv* **2023**, arXiv:2305.02765.
426. Kumar, R.; Seetharaman, P.; Luebs, A.; Kumar, I.; Kumar, K. High-fidelity audio compression with improved rvqgan. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 9–15 December 2024; pp. 1–14.
427. Wang, Y.; Wang, X.; Zhu, P.; Wu, J.; Li, H.; Xue, H.; Zhang, Y.; Xie, L.; Bi, M. Opencpop: A high-quality open source chinese popular song corpus for singing voice synthesis. *arXiv* **2022**, arXiv:2201.07429.
428. Zhang, L.; Li, R.; Wang, S.; Deng, L.; Liu, J.; Ren, Y.; He, J.; Huang, R.; Zhu, J.; Chen, X.; et al. M4singer: A multi-style, multi-singer and musical score provided mandarin singing corpus. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 6914–6926.
429. Shi, J.; Lin, Y.; Bai, X.; Zhang, K.; Wu, Y.; Tang, Y.; Yu, Y.; Jin, Q.; Watanabe, S. Singing Voice Data Scaling-up: An Introduction to ACE-Opencpop and KiSing-v2. *arXiv* **2024**, arXiv:2401.17619.
430. Timedomain. ACE Studio. Available online: https://acestudio.ai/ (accessed on 15 June 2024).
431. Ofuton-P. Available online: https://sites.google.com/view/oftn-utagoedb/%E3%83%9B%E3%83%BC%E3%83%A0 (accessed on 15 June 2024).
432. Oniku Kurumi. Available online: https://onikuru.info/db-download/ (accessed on 15 June 2024).
433. Ogawa, I. and Morise, M. Tohoku Kiritan singing database: A singing database for statistical parametric singing synthesis using Japanese pop songs. *Acoust. Sci. Technol.* **2021**, *42*, 140–145. [CrossRef]
434. Tamaru, H.; Takamichi, S.; Tanji, N.; Saruwatari, H. JVS-MuSiC: Japanese multispeaker singing-voice corpus. *arXiv* **2020**, arXiv:2001.07044.
435. Lu, P.; Wu, J.; Luan, J.; Tan, X.; Zhou, L. XiaoiceSing: A high-quality and integrated singing voice synthesis system. *arXiv* **2020**, arXiv:2006.06261.

436. Zhang, Y.; Cong, J.; Xue, H.; Xie, L.; Zhu, P.; Bi, M. Visinger: Variational inference with adversarial learning for end-to-end singing voice synthesis. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Virtual, 22–27 May 2022; pp. 7237–7241.
437. Zhang, Y.; Xue, H.; Li, H.; Xie, L.; Guo, T.; Zhang, R.; Gong, C. Visinger 2: High-fidelity end-to-end singing voice synthesis enhanced by digital signal processing synthesizer. *arXiv* **2022**, arXiv:2211.02903.
438. Yamamoto, R.; Yoneyama, R.; Toda, T. Nnsvs: A neural network-based singing voice synthesis toolkit. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; pp. 1–5.
439. Shi, J.; Guo, S.; Huo, N.; Zhang, Y.; Jin, Q. Sequence-to-sequence singing voice synthesis with perceptual entropy loss. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Virtual, 6–11 June 2021; pp. 76–80.
440. Yamamoto, R.; Yoneyama, R.; Violeta, L.P.; Huang, W.C.; Toda, T. A Comparative Study of Voice Conversion Models with Large-Scale Speech and Singing Data: The T13 Systems for the Singing Voice Conversion Challenge 2023. In Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Taipei, Taiwan, 16–20 December 2023; pp. 1–6.
441. Chen, S.; Wang, C.; Chen, Z.; Wu, Y.; Liu, S.; Chen, Z.; Li, J.; Kanda, N.; Yoshioka, T.; Xiao, X. and Wu, J. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE J. Sel. Top. Signal Process.* **2022**, *16*, 1505–1518. [CrossRef]
442. Qian, K.; Zhang, Y.; Gao, H.; Ni, J.; Lai, C.I.; Cox, D.; Hasegawa-Johnson, M.; Chang, S. Contentvec: An improved self-supervised speech representation by disentangling speakers. In Proceedings of the International Conference on Machine Learning, Baltimore, MD, USA, 17–23 July 2022; pp. 18003–18017.
443. Shi, J.; Inaguma, H.; Ma, X.; Kulikov, I.; Sun, A. Multi-resolution HuBERT: Multi-resolution Speech Self-Supervised Learning with Masked Unit Prediction. In Proceedings of the International Conference on Learning Representations (ICLR), Kigali, Rwanda, 1–5 May 2023; pp. 1–37.
444. Chen, W.; Shi, J.; Yan, B.; Berrebbi, D.; Zhang, W.; Peng, Y.; Chang, X.; Maiti, S.; Watanabe, S. Joint Prediction and Denoising for Large-Scale Multilingual Self-Supervised Learning. In Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Taipei, Taiwan, 16–20 December 2023; pp. 1–8.
445. Hsu, W.N.; Bolte, B.; Tsai, Y.H.H.; Lakhotia, K.; Salakhutdinov, R.; Mohamed, A. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 3451–3460. [CrossRef]
446. Liu, H.; Chen, Z.; Yuan, Y.; Mei, X.; Liu, X.; Mandic, D.; Wang, W.; Plumbley, M.D. AudioLDM: Text-to-Audio Generation with Latent Diffusion Models. In Proceedings of the International Conference on Machine Learning, Honolulu, HI, USA, 23–29 July 2023; pp. 21450–21474.
447. Liu, H.; Yuan, Y.; Liu, X.; Mei, X.; Kong, Q.; Tian, Q.; Wang, Y.; Wang, W.; Wang, Y.; Plumbley, M.D. Audioldm 2: Learning holistic audio generation with self-supervised pretraining. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2024**, *32*, 2871–2883. [CrossRef]
448. Liu, H.; Chen, K.; Tian, Q.; Wang, W.; Plumbley, M.D. AudioSR: Versatile audio super-resolution at scale. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Republic of Korea, 14–19 April 2024; pp. 1076–1080.
449. Loizou, P.C. *Speech Enhancement: Theory and Practice*; CRC Press: Boca Raton, FL, USA, 2007.
450. Traditional Speech Enhancement. Available online: https://github.com/fchest/traditional-speech-enhancement (accessed on 16 June 2024).
451. Hao, X.; Su, X.; Horaud, R.; Li, X. Fullsubnet: A full-band and sub-band fusion model for real-time single-channel speech enhancement. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Virtual, 6–11 June 2021; pp. 6633–6637.
452. Yu, P.; Xia, Z.; Fei, J.; Lu, Y. A survey on deepfake video detection. *IET Biom.* **2021**, *10*, 607–624. [CrossRef]
453. Zheng, X.; Guo, Y.; Huang, H.; Li, Y.; He, R. A survey of deep facial attribute analysis. *Int. J. Comput. Vis.* **2020**, *128*, 2002–2034. [CrossRef]
454. Rana, M.S.; Nobi, M.N.; Murali, B.; Sung, A.H. Deepfake detection: A systematic literature review. *IEEE Access* **2022**, *10*, 25494–25513. [CrossRef]
455. Tolosana, R.; Vera-Rodriguez, R.; Fierrez, J.; Morales, A.; Ortega-Garcia, J. Deepfakes and beyond: A survey of face manipulation and fake detection. *Inf. Fusion* **2020**, *64*, 131–148. [CrossRef]
456. Khodabakhsh, A.; Akhtar, Z. Unknown presentation attack detection against rational attackers. *IET Biom.* **2021**, *10*, 1–20. [CrossRef]
457. Malik, A.; Kuribayashi, M.; Abdullahi, S.M.; Khan, A.N. DeepFake detection for human face images and videos: A survey. *IEEE Access* **2022**, *10*, 18757–18775. [CrossRef]
458. Lyu, S. Deepfake detection: Current challenges and next steps. In Proceedings of the 2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), London, UK, 6–10 July 2020; pp. 1–6.
459. Kaddar, B.; Fezza, S.; Hamidouche,W.; Akhtar, Z.; Hadid, A. HCiT: Deepfake Video Detection Using a Hybrid Model of CNN features and Vision Transformer. In Proceedings of the 2021 IEEE Visual Communications and Image Processing (VCIP), Munich, Germany, 5–10 December 2021; pp. 1–5.
460. Heidari, A.; Jafari Navimipour, N.; Dag, H.; Unal, M. Deepfake detection using deep learning methods: A systematic and comprehensive review. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2023**, *14*, e1520. [CrossRef]
461. Khanjani, Z.; Watson, G.; Janeja, V.P. Audio deepfakes: A survey. *Front. Big Data* **2023**, *5*, 1001063. [CrossRef] [PubMed]

462. Yavuzkilic, S.; Akhtar, Z.; Sengur, A.; Siddique, K. DeepFake Face Video Detection using Hybrid Deep Residual Networks and LSTM Architecture. In *AI and Deep Learning in Biometric Security: Trends, Potential and Challenges*; CRC Press: Boca Raton, FL, USA, 2021; pp. 81–104.
463. Salman, S.; Shamsi, J.A.; Qureshi, R. Deep Fake Generation and Detection: Issues, Challenges, and Solutions. *IT Prof.* **2023**, *25*, 52–59. [CrossRef]
464. Nguyen, T.T.; Nguyen, Q.V.; Nguyen, D.T.; Nguyen, D.T.; Huynh-The, T.; Nahavandi, S.; Nguyen, T.T.; Pham, Q.V.; Nguyen, C.M. Deep learning for deepfakes creation and detection: A survey. *Comput. Vis. Image Underst.* **2022**, *223*, 103525. [CrossRef]
465. Bekci, B.; Akhtar, Z.; Ekenel, H.K. Cross-Dataset Face Manipulation Detection. In Proceedings of the 28th Signal Processing and Communications Applications Conference (SIU), Gaziantep, Turkiye, 5–7 October 2020; pp. 1–4.
466. Firc, A.; Malinka, K.; Hanáček, P. Deepfakes as a threat to a speaker and facial recognition: An overview of tools and attack vectors. *Heliyon* **2023**, *9*, e15090. [CrossRef]
467. Wu, H.; Hui, P.; Zhou, P. Deepfake in the Metaverse: An Outlook Survey. *arXiv* **2023**, arXiv:2306.07011.
468. Deng, J.; Lin, C.; Hu, P.; Shen, C.; Wang, Q.; Li, Q.; Li, Q. Towards benchmarking and evaluating deepfake detection. *IEEE Trans. Dependable Secur. Comput.* **2024**, *14*, 1–16. [CrossRef]
469. Wang, L.; Meng, X.; Li, D.; Zhang, X.; Ji, S.; Guo, S. DEEPFAKER: A unified evaluation platform for facial deepfake and detection models. *ACM Trans. Priv. Secur.* **2024**, *27*, 1–34. [CrossRef]
470. Yan, Z.; Zhang, Y.; Yuan, X.; Lyu, S.; Wu, B. Deepfakebench: A comprehensive benchmark of deepfake detection. *arXiv* **2023**, arXiv:2307.01426.
471. Lu, Y.; Ebrahimi, T. Assessment framework for deepfake detection in real-world situations. *Eurasip J. Image Video Process.* **2024**, *2024*, 6. [CrossRef]
472. Kaddar, B.; Fezza, S.A.; Akhtar, Z.; Hamidouche, W.; Hadid, A.; Serra-Sagristà, J. Deepfake Detection Using Spatiotemporal Transformer. *ACM Trans. Multimed. Comput. Commun. Appl.* **2024**, 1551–6857. [CrossRef]
473. Chen, T.; Kumar, A.; Nagarsheth, P.; Sivaraman, G.; Khoury, E. Generalization of Audio Deepfake Detection. In Proceedings of the Speaker and Language Recognition Workshop (Odyssey 2020), Tokyo, Japan, 1–5 November 2020; pp. 132–137.
474. Liu, B.; Liu, B.; Ding, M.; Zhu, T.; Yu, X. TI2Net: Temporal Identity Inconsistency Network for Deepfake Detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–7 January 2023; pp. 4691–4700.
475. Huang, B.; Wang, Z.; Yang, J.; Ai, J.; Zou, Q.; Wang, Q.; Ye, D. Implicit Identity Driven Deepfake Face Swapping Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 4490–4499.
476. Hanifa, R.M.; Isa, K.; Mohamad, S. A review on speaker recognition: Technology and challenges. *Comput. Electr. Eng.* **2021**, *90*, 107005. [CrossRef]
477. Raza, M.A.; Malik, K.M. Multimodaltrace: Deepfake detection using audiovisual representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 993–1000.
478. Salvi, D.; Liu, H.; Mandelli, S.; Bestagini, P.; Zhou, W.; Zhang, W.; Tubaro, S. A robust approach to multimodal deepfake detection. *J. Imaging* **2023**, *9*, 122. [CrossRef] [PubMed]
479. Kong, C.; Chen, B.; Yang, W.; Li, H.; Chen, P.; Wang, S. Appearance Matters, So Does Audio: Revealing the Hidden Face via Cross-Modality Transfer. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 423–436. [CrossRef]
480. Zou, H.; Shen, M.; Hu, Y.; Chen, C.; Chng, E.S.; Rajan, D. Cross-Modality and Within-Modality Regularization for Audio-Visual DeepFake Detection. *arXiv* **2024**, arXiv:2401.05746.
481. Aliev, A.; Iskakov, K. Avatarify Python. Available online: https://github.com/alievk/avatarify-python (accessed on 31 March 2024).
482. Content Authenticity Initiative (CAI). Available online: https://contentauthenticity.org/ (accessed on 30 March 2024).
483. Coalition for Content Provenance and Authority (C2PA). Available online: https://c2pa.org/ (accessed on 30 March 2024).
484. Wust, K.; Gervais, A. Do you need a blockchain? In Proceedings of the IEEE Crypto Valley Conference on Blockchain Technology (CVCBT), Zug, Switzerland, 20–22 June 2018; pp. 45–54.
485. Monrat, A.A.; Schelén, O.; Andersson, K. A survey of blockchain from the perspectives of applications, challenges, and opportunities. *IEEE Access* **2019**, *7*, 117134–117151. [CrossRef]
486. Bhutta, M.N.M.; Khwaja, A.A.; Nadeem, A.; Ahmad, H.F.; Khan, M.K.; Hanif, M.A.; Song, H.; Alshamari, M.; Cao, Y. A survey on blockchain technology: Evolution, architecture and security. *IEEE Access* **2021**, *9*, 61048–61073. [CrossRef]
487. Guo, H.; Yu, X. A survey on blockchain technology and its security. *Blockchain Res. Appl.* **2022**, *3*, 100067. [CrossRef]
488. Haut, K.; Wohn, C.; Antony, V.; Goldfarb, A.; Welsh, M.; Sumanthiran, D.; Jang, J.Z.; Ali, M.R.; Hoque, E. Could you become more credible by being White? Assessing impact of race on credibility with deepfakes. *arXiv* **2021**, arXiv:2102.08054.
489. Lea, C.; Flynn, M.D.; Vidal, R.; Reiter, A.; Hager, G.D. Temporal convolutional networks for action segmentation and detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 156–165.
490. Ghojogh, B.; Ghodsi, A.; Karray, F.; Crowley, M. Generative adversarial networks and adversarial autoencoders: Tutorial and survey. *arXiv* **2021**, arXiv:2111.13282.
491. Huang, X.; Li, Y.; Poursaeed, O.; Hopcroft, J.; Belongie, S. Stacked generative adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5077–5086.

492. Zhang, H.; Xu, T.; Li, H.; Zhang, S.; Wang, X.; Huang, X.; Metaxas, D.N. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 1947–1962. [CrossRef] [PubMed]
493. Zhang, H.; Xu, T.; Li, H.; Zhang, S.; Wang, X.; Huang, X.; Metaxas, D.N. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5907–5915.
494. Guo, Z.; Wang, L.; Yang, W.; Yang, G.; Li, K. LDFNet: Lightweight dynamic fusion network for face forgery detection by integrating local artifacts and global texture information. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *34*, 1255–1265. [CrossRef]
495. Wang, L.Y.; Akhtar, Z. CCAP: Cooperative Context Aware Pruning for Neural Network Model Compression. In Proceedings of the IEEE International Symposium on Multimedia (ISM), Naple, Italy, 29 November–1 December 2021; pp. 257–260.
496. Feldmann, J.; Youngblood, N.; Karpov, M.; Gehring, H.; Li, X.; Stappers, M.; Le Gallo, M.; Fu, X.; Lukashchuk, A.; Raja, A.S.; et al. Parallel convolutional processing using an integrated photonic tensor core. *Nature* **2021**, *589*, 52–58. [CrossRef] [PubMed]
497. Muñoz, A.; Rios, R.; Román, R.; López, J. A survey on the (in) security of trusted execution environments. *Comput. Secur.* **2023**, *129*, 103180. [CrossRef]
498. Kaddar, B.; Fezza, S.A.; Hamidouche, W.; Akhtar, Z.; Hadid, A. On the effectiveness of handcrafted features for deepfake video detection. *J. Electron. Imaging* **2023**, *32*, 053033. [CrossRef]
499. Bin, Z.; Zhao, H.; Liang, X.; Chen, W. FSA-Net: A Cost-efficient Face Swapping Attention Network with Occlusion-Aware Normalization. *Intell. Autom. Soft Comput.* **2023**, *37*, 971–983. [CrossRef]
500. Mittal, G.; Jakobsson, A.; Marshall, K.O.; Hegde, C.; Memon, N. AI-assisted Tagging of Deepfake Audio Calls using Challenge-Response. *arXiv* **2024**, arXiv:2402.18085.
501. Korshunov, P.; Marcel, S. Deepfake detection: Humans vs. machines. *arXiv* **2020**, arXiv:2009.03155.
502. Müller, N.M.; Pizzi, K.; Williams, J. Human perception of audio deepfakes. In Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia, Lisboa, Portugal, 14 October 2022; pp. 85–91.
503. Nistal, J.; Lattner, S.; Richard, G. Comparing representations for audio synthesis using generative adversarial networks. In Proceedings of the 2020 28th European Signal Processing Conference (EUSIPCO), Amsterdam, The Netherlands, 18–22 January 2021; pp. 161–165.
504. Theis, L.; Oord, A.V.D.; Bethge, M. A note on the evaluation of generative models. *arXiv* **2015**, arXiv:1511.01844.
505. Croitoru, F.A.; Hondru, V.; Ionescu, R.T.; Shah, M. Diffusion models in vision: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 10850–10869. [CrossRef] [PubMed]
506. Yang, L.; Zhang, Z.; Song, Y.; Hong, S.; Xu, R.; Zhao, Y.; Zhang, W.; Cui, B.; Yang, M.H. Diffusion models: A comprehensive survey of methods and applications. *ACM Comput. Surv.* **2023**, *56*, 1–39. [CrossRef]
507. Wang, Y.;Su, Z.; Zhang, N.; Xing, R.; Liu, D.; Luan, T.H.; Shen, X. A survey on metaverse: Fundamentals, security, and privacy. *IEEE Commun. Surv. Tutor.* **2022**, *25*, 319–352. [CrossRef]
508. Stavola, J.; Choi, K.S. Victimization by Deepfake in the Metaverse: Building a Practical Management Framework. *Int. J. Cybersecur. Intell. Cybercrime* **2023**, *6*, 2. [CrossRef]
509. International Organization for Standardization (ISO). Available online: https://www.iso.org/foresight/computing.html (accessed on 31 March 2024).
510. Mueck, M.; Forbes, R.; Cadzow, S.; Wood, S.; Gazis, E. ETSI Activities in the Field of Artificial Intelligence—Preparing the Implementation of the European AI Act. ETSI, Sophia Antipolis, France, White Paper, 52. 2022. Available online: https://www.etsi.org/newsroom/press-releases/2167-etsi-s-activities-in-artificial-intelligence-read-our-new-white-paper (accessed on 31 March 2024).
511. Diakopoulos, N.; Johnson, D. Anticipating and addressing the ethical implications of deepfakes in the context of elections. *New Media Soc.* **2021**, *23*, 2072–2098. [CrossRef]
512. Pantserev, K.A. The malicious use of AI-based deepfake technology as the new threat to psychological security and political stability. In *Cyber Defence in the Age of AI, Smart Societies and Augmented Humanity*; Springer: Cham, Switzerland, 2020; pp. 37–55.
513. Zhou, J.; Pun, C.M. Personal privacy protection via irrelevant faces tracking and pixelation in video live streaming. *IEEE Trans. Inf. Forensics Secur.* **2020**, *16*, 1088–1103. [CrossRef]
514. Wang, Y.; Chen, S.; Yao, T.; Ma, L.; Zhang, Z.; Tan, X. Explore and Enhance the Generalization of Anomaly DeepFake Detection. In Proceedings of the International Conference on Computational Visual Media, Wellington, New Zealand, 10–12 April 2024; pp. 27–47.
515. Yang, Z.; Liang, J.; Xu, Y.; Zhang, X.Y.; He, R. Masked relation learning for deepfake detection. *IEEE Trans. Inf. Forensics Secur.* **2023**, *18*, 1696–1708. [CrossRef]
516. Tang, Y. ECGAN: Translate Real World to Cartoon Style Using Enhanced Cartoon Generative Adversarial Network. *Comput. Mater. Contin.* **2023**, *76*, 1195–1212. [CrossRef]
517. Wu, F.; Ma, Y.; Zhang, Z. I Found a More Attractive Deepfaked Self: The Self-Enhancement Effect in Deepfake Video Exposure. *Cyberpsychol. Behav. Soc. Netw.* **2021**, *24*, 173–181. [CrossRef]
518. Zhao, Y.; Liu, B.; Ding, M.; Liu, B.; Zhu, T.; Yu, X. Proactive deepfake defence via identity watermarking. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA 3–7 January 2023; pp. 4602–4611.
519. Sun, P.; Li, Y.; Qi, H.; Lyu, S. Landmark breaker: Obstructing deepfake by disturbing landmark extraction. In Proceedings of the 2020 IEEE International Workshop on Information Forensics and Security (WIFS), Virtual, 6–9 December 2020; pp. 1–6.

520. Bao, J.; Chen, D.; Wen, F.; Li, H.; Hua, G. Towards open-set identity preserving face synthesis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6713–6722.
521. Xia, W.; Yang, Y.; Xue, J.H.; Wu, B. Tedigan: Text-guided diverse face image generation and manipulation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 2256–2265.
522. Ferraro, M.F. Deepfake Legislation: A Nationwide Survey—State and Federal Lawmakers Consider Legislation to Regulate Manipulated Media. WilmerHale Report: Deepfake Legislation: A Nationwide Survey—State and Federal Lawmakers Consider Legislation to Regulate Manipulated Media. 2019. Available online: https://www.wilmerhale.com/insights/client-alerts/20190925-deepfake-legislation-a-nationwide-survey (accessed on 31 March 2024).
523. Dixit, A.; Kaur, N.; Kingra, S. Review of audio deepfake detection techniques: Issues and prospects. *Expert Syst.* **2023**, *40*, e13322. [CrossRef]
524. Shih, T.H.; Yeh, C.Y.; Chen, M.S. Does Audio Deepfake Detection Rely on Artifacts? In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Republic of Korea, 14–19 April 2024; pp. 12446–12450.
525. Ballesteros, D.M.; Rodriguez-Ortega, Y.; Renza, D.; Arce, G. Deep4SNet: Deep learning for fake speech classification. *Expert Syst. Appl.* **2021**, *184*, 115465. [CrossRef]
526. Wang, R.; Huang, Z.; Chen, Z.; Liu, L.; Chen, J.; Wang, L. Anti-forgery: Towards a stealthy and robust deepfake disruption attack via adversarial perceptual-aware perturbations. *arXiv* **2022**, arXiv:2206.00477.
527. Sablayrolles, A.; Douze, M.; Schmid, C.; Jégou, H. Radioactive data: Tracing through training. In Proceedings of the International Conference on Machine Learning, Virtual, 12–18 July 2020; pp. 8326–8335.
528. Adadi, A.; Berrada, M. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access* **2018**, *6*, 52138–52160. [CrossRef]
529. Das, A.; Rad, P. Opportunities and challenges in explainable artificial intelligence (xai): A survey. *arXiv* **2020**, arXiv:2006.11371.
530. Linardatos, P.; Papastefanopoulos, V.; Kotsiantis, S. Explainable ai: A review of machine learning interpretability methods. *Entropy* **2020**, *23*, 18. [CrossRef]
531. Arrieta, A.B.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; García, S.; Gil-López, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **2020**, *58*, 82–115. [CrossRef]
532. Schwalbe, G.; Finzel, B. A comprehensive taxonomy for explainable artificial intelligence: A systematic survey of surveys on methods and concepts. *Data Min. Knowl. Discov.* **2023**, 1–59. [CrossRef]
533. Zhang, Y.; Yang, Q. An overview of multi-task learning. *Natl. Sci. Rev.* **2018**, *5*, 30–43. [CrossRef]
534. Tarasiou, M.; Zafeiriou, S. Extracting Deep Local Features to Detect Manipulated Images of Human Faces. In Proceedings of the 2020 IEEE International Conference on Image Processing (ICIP), Abu Dhabi, UAE, 25–28 September 2020; pp. 1821–1825.
535. Kadam, S.; Vaidya, V. Review and analysis of zero, one and few shot learning approaches. In Proceedings of the Intelligent Systems Design and Applications: 18th International Conference on Intelligent Systems Design and Applications (ISDA), Vellore, India, 6–8 December 2018; pp. 100–112.
536. Song, Y.; Wang, T.; Cai, P.; Mondal, S.K.; Sahoo, J.P. A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities. *ACM Comput. Surv.* **2023**, *55*, 1–40. [CrossRef]
537. Groh, M.; Epstein, Z.; Firestone, C.; Picard, R. Deepfake detection by human crowds, machines, and machine-informed crowds. *Proc. Natl. Acad. Sci. USA* **2022**, *119*, e2110013119. [CrossRef]
538. Bray, S.D.; Johnson, S.D.; Kleinberg, B. Testing human ability to detect 'deepfake' images of human faces. *J. Cybersecur.* **2023**, *9*, tyad011. [CrossRef]
539. Ding, F.; Zhu, G.; Li, Y.; Zhang, X.; Atrey, P.K.; Lyu, S. Anti-forensics for face swapping videos via adversarial training. *IEEE Trans. Multimed.* **2021**, *24*, 3429–3441. [CrossRef]
540. Monteiro, J.; Albuquerque, I.; Akhtar, Z.; Falk, T.H. Generalizable adversarial examples detection based on bi-model decision mismatch. In Proceedings of the 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC), Bari, Italy, 6–9 October 2019; pp. 2839–2844.
541. Kabir, M.M.; Mridha, M.F.; Shin, J.; Jahan, I.; Ohi, A.Q. A survey of speaker recognition: Fundamental theories, recognition methods and opportunities. *IEEE Access* **2021**, *9*, 79236–79263. [CrossRef]
542. Ohi, A.Q.; Mridha, M.F.; Hamid, M.A.; Monowar, M.M. Deep speaker recognition: Process, progress, and challenges. *IEEE Access* **2021**, *9*, 89619–89643. [CrossRef]
543. Sadjadi, S.O.; Greenberg, C.; Singer, E.; Mason, L.; Reynolds, D. The 2021 NIST speaker recognition evaluation. *arXiv* **2022**, arXiv:2204.10242.
544. Huh, J.; Brown, A.; Jung, J.W.; Chung, J.S.; Nagrani, A.; Garcia-Romero, D.; Zisserman, A. Voxsrc 2022: The fourth voxceleb speaker recognition challenge. *arXiv* **2023**, arXiv:2302.10248.
545. Dash, B.; Sharma, P. Are ChatGPT and deepfake algorithms endangering the cybersecurity industry? A review. *Int. J. Eng. Appl. Sci.* **2023**, *10*, 21–39.
546. Peng, R.D.; Hicks, S.C. Reproducible research: A retrospective. *Annu. Rev. Public Health* **2021**, *42*, 79–93. [CrossRef]
547. Tampubolon, M. Digital Face Forgery and the Role of Digital Forensics. *Int. J. Semiot.-Law-Rev. Int. SéMiotique Jurid.* **2023**, *37*, 1–15. [CrossRef]

548. Mcuba, M.; Singh, A.; Ikuesan, R.A.; Venter, H. The effect of deep learning methods on deepfake audio detection for digital investigation. *Procedia Comput. Sci.* **2023**, *219*, 211–219. [CrossRef]
549. Newman, L.H. Police Bodycams Can Be Hacked to Doctor Footage. 2018. Available online: https://www.wired.com/story/police-body-camera-vulnerabilities/ (accessed on 25 December 2023).
550. Open Media Forensics Challenge. Available online: https://mfc.nist.gov/ (accessed on 20 December 2023).
551. Iproov. Available online: https://www.iproov.com/blog/deepfakes-statistics-solutions-biometric-protection (accessed on 25 December 2023).
552. Helmus, T.C. Artificial Intelligence, Deepfakes, and Disinformation. 2022, pp. 1–24. Available online: https://www.rand.org/content/dam/rand/pubs/perspectives/PEA1000/PEA1043-1/RAND_PEA1043-1.pdf (accessed on 4 April 2024).
553. FotoForensics. Available online: https://fotoforensics.com/ (accessed on 4 April 2024).
554. InVID Project—Video Verification. Available online: https://www.invid-project.eu/tools-and-services/invid-verification-plugin/ (accessed on 4 April 2024).
555. WeVerify. Available online: https://weverify.eu/ (accessed on 4 April 2024).
556. Godulla, A.; Hoffmann, C.P.; Seibert, D. Dealing with deepfakes–an interdisciplinary examination of the state of research and implications for communication studies. *SCM Stud. Commun. Media* **2021**, *10*, 72–96. [CrossRef]
557. Brooks, C.F. Popular discourse around deepfakes and the interdisciplinary challenge of fake video distribution. *Cyberpsychol. Behav. Soc. Netw.* **2021**, *24*, 159–163. [CrossRef] [PubMed]
558. Whittaker, L.; Mulcahy, R.; Letheren, K.; Kietzmann, J.; Russell-Bennett, R. Mapping the deepfake landscape for innovation: A multidisciplinary systematic review and future research agenda. *Technovation* **2023**, *125*, 102784. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.