computers

MDPI

Review

Deepfake Attacks: Generation, Detection, Datasets, Challenges, and Research Directions

Amal Naitali [1,*], Mohammed Ridouani [1], Fatima Salahdine [2] and Naima Kaabouch [3,*]

[1] RITM Laboratory, CED Engineering Sciences, Hassan II University, Casablanca 20000, Morocco; mohammed.ridouani@etu.univh2c.ma
[2] Department of Electrical and Computer Engineering, The University of North Carolina at Charlotte, Charlotte, NC 28223, USA; fsalahdi@uncc.edu
[3] School of Electrical Engineering and Computer Science, University of North Dakota, Grand Forks, ND 58202, USA
* Correspondence: amal.naitali-etu@etu.univh2c.ma (A.N.); naima.kaabouch@und.edu (N.K.)

Abstract: Recent years have seen a substantial increase in interest in deepfakes, a fast-developing field at the nexus of artificial intelligence and multimedia. These artificial media creations, made possible by deep learning algorithms, allow for the manipulation and creation of digital content that is extremely realistic and challenging to identify from authentic content. Deepfakes can be used for entertainment, education, and research; however, they pose a range of significant problems across various domains, such as misinformation, political manipulation, propaganda, reputational damage, and fraud. This survey paper provides a general understanding of deepfakes and their creation; it also presents an overview of state-of-the-art detection techniques, existing datasets curated for deepfake research, as well as associated challenges and future research trends. By synthesizing existing knowledge and research, this survey aims to facilitate further advancements in deepfake detection and mitigation strategies, ultimately fostering a safer and more trustworthy digital environment.

Keywords: deepfake detection; face forgery; deep learning; generative artificial intelligence; vision transformers

Citation: Naitali, A.; Ridouani, M.; Salahdine, F.; Kaabouch, N. Deepfake Attacks: Generation, Detection, Datasets, Challenges, and Research Directions. *Computers* **2023**, *12*, 216. https://doi.org/10.3390/computers12100216

Academic Editors: Aditya Kumar Sahu, Amine Khaldi and Jatindra Kumar Dash

Received: 19 September 2023
Revised: 11 October 2023
Accepted: 20 October 2023
Published: 23 October 2023

1. Introduction

Deepfakes are produced by manipulating existing videos and images to produce realistic-looking but wholly fake content. The rise of advanced artificial intelligence-based tools and software that require no technical expertise has made deepfake creation easier. With the unprecedented exponential advancement, the world is currently witnessing in generative artificial intelligence, the research community is in dire need of keeping informed on the most recent developments in deepfake generation and detection technologies to not fall behind in this critical arms race.

Deepfakes present a number of serious issues that arise in a variety of fields. These issues could significantly impact people, society [1], and the reliability of digital media [2]. Some significant issues include fake news, which can lead to the propagation of deceptive information, manipulation of public opinion, and erosion of trust in media sources. Deepfakes can also be employed as tools for political manipulation, influence elections, and destabilize public trust in political institutions [3,4]. In addition, this technology enables malicious actors to create and distribute non-consensual explicit content to harass and cause reputational damage or create convincing impersonations of individuals, deceiving others for financial or personal gains [5]. Furthermore, the rise of deepfakes poses a serious issue in the domain of digital forensics as it contributes to a general crisis of trust and authenticity in digital evidence used in litigation and criminal justice proceedings. All of these impacts show that deepfakes present a serious threat, especially in the current sensitive state of the international political climate and the high stakes at hand considering

the conflicts on the global scene and how deepfakes and fake news can be weaponized in the ongoing media war, which can ultimately result in catastrophic consequences.

Therefore, deepfake detection techniques need to be constantly improved to catch up with the fast-paced evolution of generative artificial intelligence. There is a need for literature reviews to keep up with the fast-changing field of artificial intelligence and deepfakes to enable researchers and professionals to develop robust countermeasure methods and to lay the right groundwork to make it easier to detect and mitigate deepfakes.

The key contributions to this survey paper are as follows:

- A summary of the state-of-the art deepfake generation and detection techniques;
- An overview of fundamental deep learning architectures used as backbone in deepfake video detection models;
- A list of existing deepfake datasets contributing to the improvement of the performance, generalization and robustness of deepfake detection models;
- A discussion of the limitations of existing techniques, challenges, and research directions in the field of deepfake detection and mitigation.

The remainder of this paper is organized as follows. Section 2 provides an outline of the most recent, existing survey papers related to deepfake technology. Section 3 is devoted to deepfake manipulation techniques for generating deepfakes. Section 4 describes existing deepfake detection techniques. Section 5 gives a list of existing datasets used for deepfake research. In Section 6, we discuss some of the challenges and future research directions of the deepfake field. Finally, the survey ends with a conclusion.

2. Related Surveys

Multiple surveys of the literature in the area of deepfake detection have been published in recent years as the topic is advancing rapidly. For instance, the authors of [6] offered a systematic literature review with a new, interdisciplinary viewpoint on deepfakes. They provided a meticulous definition of deepfakes and discussed the impact of the creation and spread of deepfakes. They also suggested future research directions for innovation. Alternatively, the authors of [7] provided a rich review paper that has an exhaustive breakdown of deepfake types alongside the technology leveraged in their creation and detection, as well as open-source datasets and future trends in deepfake technology. In Ref. [8], the authors focused in their systematic review on deepfake detection-technology. They include machine learning and deep learning methods alongside statistical techniques and blockchain-based techniques, assessed how well each method performs when applied to diverse datasets, and offered some recommendations on deepfake detection that may aid future studies. In Ref. [9], the author presented recent deepfake advancements, covering four face manipulation types—generation, detection methods, and future prospects.

In Ref. [10], the authors explored the background and methods of deepfakes before looking at the development of improved and resilient deep learning techniques to combat their use. In Ref. [11], the authors provided a survey with an extensive summary of deepfake manipulation types, the tools and technology used to generate deepfakes, a technical background of deep learning models and public datasets, and the challenges faced in the creation and detection of deepfakes. Whereas, in [12], the authors presented a detailed review of deepfake manipulation types and their generation processes, as well as several detection methods and the features leveraged, alongside some issues that demand serious consideration in future studies. The authors of [13] offered in their survey a technical background on the architecture used in deepfake creation that deals with two manipulation types: reenactment and replacement. In addition to detection technologies and prevention solutions, they mentioned several inadequacies of the available defense options and areas that need more focus.

In a detailed survey [14], the authors covered several topics of deepfake manipulation, including audio deepfakes, the technology used in its creation and detection, performance metrics, and publicly available datasets, in addition to a discussion about the limitations and future trends in the field of deepfakes. An analysis of several CNN- and RNN-based

deepfake video detection models was described in [15]. In addition, other surveys [16–20] offered a boiled-down summary of the principal elements in the field of deepfakes, such as their definition, impact, creation process, and detection methods. Table 1 gives a summary of topics covered and not covered by the above-mentioned survey papers.

Table 1. An overview of related deepfake surveys and coverage topics.

Author	Title	Covered	Not Covered
Sudhakar and Shanthi [21]	Deepfake: An Endanger to Cyber Security	Deepfake generation	Deepfake types
		Deepfake detection	Datasets
Salman et al. [22]	Deepfake Generation and Detection: Issues, Challenges, and Solutions	Audio–visual Deepfake generation	Datasets
		Deepfake detection	
Khder et al. [23]	Artificial Intelligence into Multimedia Deepfakes Creation and Detection	Deepfake types	Datasets
		Deepfake generation	
		Deepfake detection	
Kandari et al. [24]	A Comprehensive Review of Media Forensics and Deepfake Detection Technique	Forensic-based deepfake detection methods	Deepfake types
			Deepfake generation
			Datasets
Boutadjine et al. [25]	A comprehensive study on multimedia Deepfakes	Deepfake generation	Deepfake types
		Deepfake detection	
		Threats and limitations	Datasets
Mallet et al. [26]	Using Deep Learning to Detecting Deepfakes	Deepfake detection	Deepfake generation
		Datasets	Deepfake types
		Limitations	
Das et al. [15]	A Survey on Deepfake Video-Detection Techniques Using Deep Learning	Deep learning-based detection models	Deepfake types
			Deepfake generation
			Datasets
Alanazi [27]	Comparative Analysis of Deepfake Detection Techniques	Deepfake creation	Datasets
		Deepfake detection	
Xinwei et al. [28]	An Overview of Face Deep Forgery	Deepfake generation	Deepfake detection
		Deepfake types	
			Datasets
Weerawardana and Fernando [29]	Deepfakes Detection Methods: A Literature Survey	Deepfake detection	Deepfake types
		Limitations	Deepfake generation
P and Sk [30]	Deepfake Creation and Detection: A Survey	Deepfake generation	Deepfake types
		Deepfake detection	Datasets
Lin et al. [16]	A Survey of Deepfakes Generation and Detection	Deepfake types	Future trends
		Deepfake generation	
		Deepfake detection	
		Datasets	

Table 1. *Cont.*

Author	Title	Covered	Not Covered
Khichi and Kumar Yadav [18]	A Threat of Deepfakes as a Weapon on Digital Platforms and their Detection Methods	Deepfake generation	Datasets
		Deepfake detection	
		Limitations and future trends	
Chaudhary et al. [19]	A Comparative Analysis of Deepfake Techniques	Deepfake creation	Deepfake types
		Deepfake detection	
		Future directions	
			Datasets
Zhang et al. [31]	Deep Learning in Face Synthesis: A Survey on Deepfakes	Deepfake types	Datasets
		Deepfake generation	Deepfake detection
Younus and Hasan [20]	Abbreviated View of Deepfake Videos Detection Techniques	Deepfake generation	Deepfake types
		Deepfake detection	Datasets

3. Deepfake Generation

In this section, we will first state the various types of deepfake manipulations and then deliver an overview of deepfake generation techniques.

3.1. Deepfake Manipulation Types

There exist five primary types of deepfake manipulation, as shown in Figure 1. Face synthesis [32] is a manipulation type which entails creating images of a human face that does not exist in real life. In attribute manipulation [33], only the region that is relevant to the attribute is altered alone in order to change the facial appearance by removing or donning eyeglasses, retouching the skin, and even making some more significant changes, like changing the age and gender. Nevertheless, our attention is directed towards manipulations that are predominantly prevalent in video format due to their heightened engagement levels compared to image-based content. Consequently, it is more likely for people to fall victim to deepfake videos. These manipulations are designed to make it appear as though a person is doing or saying something that they did not actually do or say.

The most common manipulation types are identity swap or face swapping, face reenactment, and lip-syncing. Face swapping [34,35] is a form of manipulation that has primarily become prevalent in videos even though it can occur at the image level. It entails the substitution of one individual's face in a video, known as the source, with the face of another person, referred to as the target. In this process, the original facial features and expressions of the target subject are mapped onto the associated areas of the source subject's face, creating a seamless integration of the target's appearance into the source video. The origins of research on the subject of identity swap can be traced to the morphing method introduced in [36]. Meanwhile, face reenactment [37,38] is a manipulation technique that focuses on altering the facial expressions of a person in a video. It involves the replacement of the original facial expression of the subject, with the facial expression of another person. Lastly, we have lip-syncing [39], where the objective is to generate a target face that appears authentic and synchronizes with given text or audio inputs. Achieving accurate lip movements and facial expressions that align with the source audio necessitates the use of advanced techniques. Additionally, meticulous post-processing is crucial to ensuring that the resulting video portrays a natural and seamless facial appearance.

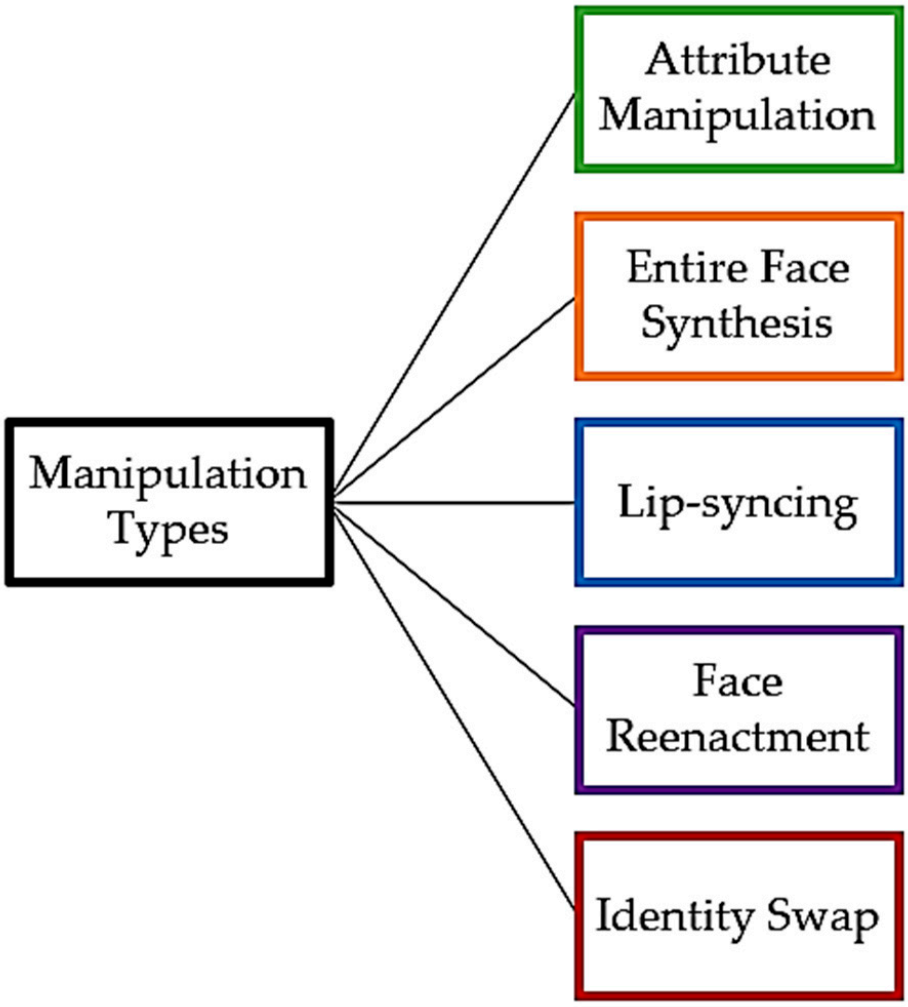

Figure 1. The five principal categories of deepfake manipulation.

3.2. Deepfake Generation Techniques

Multiple techniques exist for generating deepfakes. Generative Adversarial Networks GANs [40] and Autoencoders are the most prevalent techniques. GANs consist of a pair of neural networks, a generator network and discriminator network, which engage in a competitive process. The generator network produces synthetic images, which are presented alongside real images to the discriminator network. The generator network learns to produce images that deceive the discriminator, while the discriminator network is trained to differentiate between real and synthetic images. Through iterative training, GANs become proficient at producing increasingly realistic deepfakes. On the other hand, Autoencoders can be used as feature extractors to encode and decode facial features. During training, the autoencoder learns to compress an input facial image into a lower-dimensional representation that retains essential facial features. This latent space representation can then be used to reconstruct the original image. Though, for deepfake generation, two autoencoders are leveraged, one trained on the face of the source and another trained on the target.

Numerous sophisticated GAN-based techniques have emerged in the literature, contributing to the advancement and complexity of deepfakes. AttGAN [41] is a technology for facial attribute manipulation; its attribute awareness enables precise and high-quality attribute changes, making it valuable for applications like face-swapping and age progression or regression. Likewise, StyleGAN [42] is a GAN architecture that excels in generating highly realistic and detailed images. It allows for the manipulation of various facial features, making it a valuable tool for generating high-quality deepfakes. Similarly, STGAN [33]

modifies specific facial attributes in images while preserving the person's identity. The model can work with labeled and unlabeled data and has shown promising results in accurately controlling attribute changes. Another technique is StarGANv2 [43], which is able to perform multi-domain image-to-image translation, enabling the generation of images across multiple different domains using a single unified model. Unlike the original StarGAN [44], which could only perform one-to-one translation between each pair of domains, StarGANv2 [43] can handle multiple domains simultaneously. An additional GAN variant is CycleGAN [45], which specializes in style transfer between two domains. It can be applied to transfer facial features from one individual to another, making it useful for face-swapping applications. Moreover, there is RSGAN [46], which can encode the appearances of faces and hair into underlying latent space representations, enabling the image appearances to be modified by manipulating the representations in the latent spaces. For a given audio input, LipGAN [47] is intended to produce realistic lip motions and speech synchronization.

In addition to the previously mentioned methods, there is a range of open-source tools readily available for digital use, enabling users to create deep fakes with relative ease, like FaceApp [48], Reface [49], DeepBrain [50], DeepFaceLab [51], and Deepfakes Web [52]. These tools have captured the public's attention due to their accessibility and ability to produce convincing deepfakes. It is essential for users to utilize these tools responsibly and ethically to avoid spreading misinformation or engaging in harmful activities. As artificial intelligence is developing fast, deepfake generation algorithms are simultaneously becoming more sophisticated, convincing, and hard to detect.

4. Deepfake Detection

This section will point out the diverse clues and detection models exploited to achieve the task of classifying fake media from genuine ones. Next, it will delve into the various state-of-the-art deep learning architectures implemented in deepfake detection techniques and provide a summary of several recent deepfake detection models.

4.1. Deepfake Detection Clues

Deepfakes can be detected by exploiting various clues, as summarized in Figure 2. One approach is to analyze spatial inconsistencies by closely examining deepfakes for visual artifacts, facial landmarks, or intra-frame inconsistencies. Another method involves detecting convolutional traces that are often present in deepfakes as a result of the generation process, for instance, bi-granularity artifacts and GAN fingerprints. Additionally, biological signals such as abnormal eye blinking frequency, eye color, and heartbeat can also indicate the presence of a deepfake, as can temporal inconsistencies or the discontinuity between adjacent video frames, which may result in flickering, jittering, and changes in facial position. Poor alignment of facial emotions on swapped faces in deepfakes is a high-level semantic feature used in detection techniques. Detecting audio-visual inconsistencies is a multimodal approach that can be used for deepfakes that involve swapping both faces and audio. Another multimodal approach is to exploit spatial-temporal features by inspecting visual irregularities within individual video frames (intra-frame inspection) and analyzing temporal characteristics across video streams (inter-frame examination).

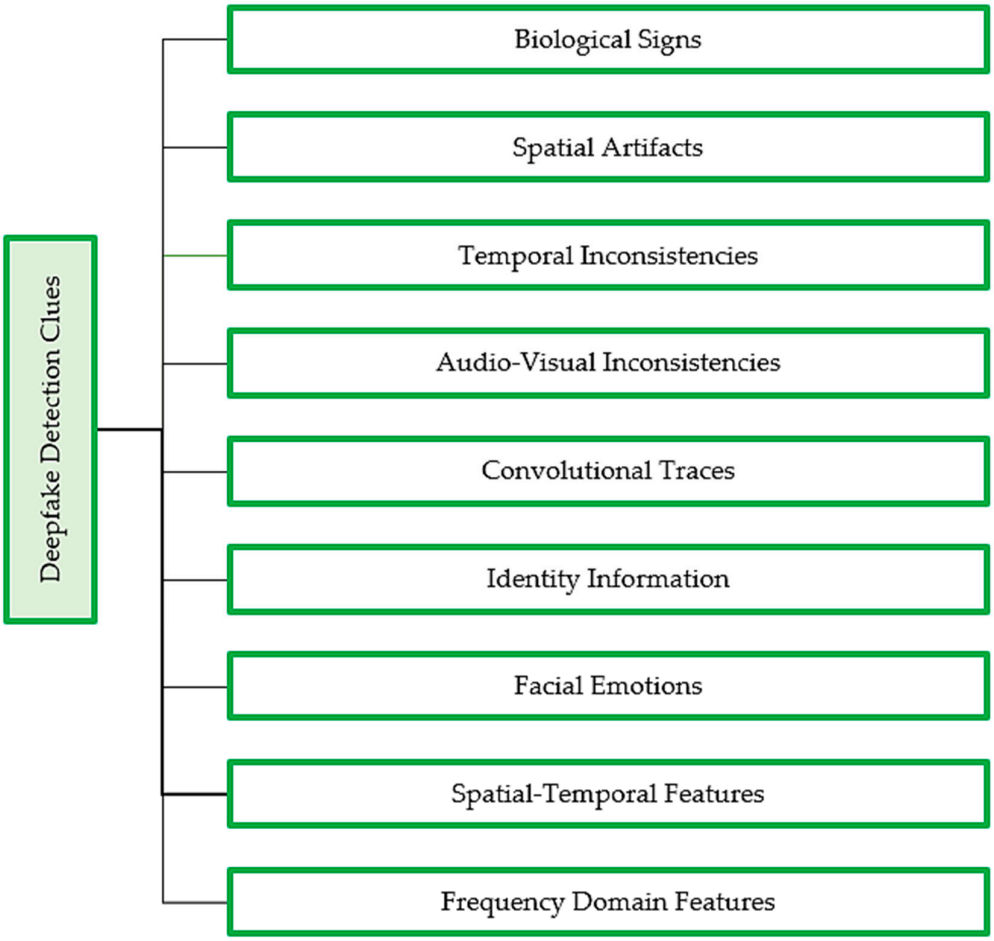

Figure 2. Clues and features employed by deepfake detection models in the identification of deepfake content.

4.1.1. Detection Based on Spatial Artifacts

To effectively use face landmark information, in Ref. [53], Liang et al. described a facial geometry prior module. The model harnesses facial maps and correlation within the frequency domain to study the distinguishing traits of altered and unmanipulated regions by employing a CNN-LSTM network. In order to predict manipulation localization, a decoder is utilized to acquire the mapping from low-resolution feature maps to pixel-level details, and SoftMax function was implemented for the classification task. A different approach, dubbed forensic symmetry, by Li, G. et al. [54], assessed whether the natural features of a pair of mirrored facial regions are identical or dissimilar. The symmetry attribute extracted from frontal facial images and the resemblance feature obtained from profiles of the face images are obtained by a multi-stream learning structure that uses DRN as its backbone network. The difference between the two symmetrical face patches is then quantified by mapping them into angular hyperspace. A heuristic prediction technique was used to put this model into functioning at the video level. As a further step, a multi-margin angular loss function was developed for classification.

Hu et al. [55] proposed DeepfakeMAE which is a detection model that can leverage the commonalities across all facial components. To be more specific, a masked autoencoder is pretrained to learn facial part consistency by randomly masking some facial features and rebuilding missing sections using the facial parts that are still visible. This is performed given a real face image. Moreover, a model employing two networks, both utilizing pre-trained encoders and decoders, is leveraged to optimize the differentiation between authentic and counterfeit videos. Yang, J. et al. [56] tackled deepfake detection from a different perspective where they simulate the fake image generation process to explore

forgery traces. A multi-scale self-texture attention Generative Network is suggested for this aim employing an encoder–decoder generator, Resnet as backbone network, and the self-texture attention method to improve the texture characteristics in the process of disassembling an image. Additionally, a loss function termed Prob-tuple loss confined by classification probability is suggested. To identify visual artifacts at different scales, Wang et al. [57] introduced a Multi-modal Multi-scale Transformer that works on patches of various sizes to identify disparities within images at various spatial tiers as well as forgery artifacts in the frequency domain; and the latter is added to RGB information by means of a cross modality fusion block. An approach based on GANs for deepfake detection is suggested by Xiao et al. [58], leveraging the concealed gradient data within the grayscale representation of the manipulated image and incorporating focal loss for the classification task.

4.1.2. Detection Based on Biological/Physiological Signs

Li, Y. et al. [59] adopted an approach based on identifying eye blinking, a biological signal that is not easily conveyed in deepfake videos. Therefore, a deepfake video can be identified by the absence of eye blinking. To spot open and closed eye states, a deep neural network model that blends CNN and a recursive neural network is used while taking into account previous temporal knowledge. Alternatively, Hernandez-Ortega et al. [60] present an innovative approach for detecting deepfake videos that focuses on analyzing heart rate information through remote photoplethysmography (rPPG). By examining video sequences and identifying slight alterations in skin color, the existence of human blood beneath the tissues can be revealed. The proposed detection system, called DeepfakesON-Phys, incorporates a Convolutional Attention Network to extract spatial and temporal details from video frames and effectively combine the two origins for improved fake video detection.

4.1.3. Detection Based on Audio-Visual Inconsistencies

Boundary Aware Temporal Forgery Detection is a multimodal technique introduced by Cai et al. [61] for correctly predicting the borders of fake segments based on visual and auditory input. While an audio encoder using a 2DCNN learns characteristics extracted from the audio, a video encoder leveraging a 3DCNN learns frame-level spatial-temporal information. Yang, W. et al. [62] also exploited discrepancy between audio and visual elements for deepfake identification. A temporal-spatial encoder for feature embedding explores the disparity between audio and visual components at temporal and spatial levels and a multi-modal joint-decoder, designed to concurrently acquire knowledge of multi-modal interactions and integrate audio-visual data, alongside the cross-modal classifier incorporated for manipulation detection. Similarly performed by considering both the audio and visual aspects of a video, Ilyas et al. [63] introduced an end-to-end method called AVFakeNet. The detection model is comprised of a Dense Swin Transformer Net (DST-Net).

4.1.4. Detection Based on Convolutional Traces

To detect deepfakes, Huang et al. [64] harnessed the imperfection of the up-sampling process in GAN-generated deepfakes by employing a map of gray-scale fakeness. Furthermore, attention mechanism, augmentation of partial data, and clustering of individual samples are employed to improve the model's robustness. Chen et al. [65] exploited a different trace which is bi-granularity artifacts, intrinsic-granularity artifacts that are caused by up-convolution or up-sampling operations, and extrinsic granularity artifacts that are the result of the post-processing step that blends the synthesized face to the original video. Deepfake detection is tackled as a multi-task learning problem where ResNet-18 is used as the backbone feature extractor. Whereas L. Guarnera et al. [66] provided a method that uses an expectation maximization algorithm to extract a set of local features intended to simulate the convolutional patterns frequently found in photos. The five currently accessible architectures are GDWCT [67], StarGAN [68], AttGAN [41], StyleGAN [42], and

StyleGAN2 [69]. Next, naive classifiers are trained to differentiate between real images and those produced by these designs.

4.1.5. Detection Based on Identity Information

Based on the intuition that every person can exhibit distinct patterns in the simultaneous occurrence of their speech, facial expressions, and gestures, Agarwal et al. [70] introduced a multimodal detection method with a semantic focus that incorporates speech transcripts into gestures specific to individuals analysis using interpretable action units to model facial and cranial motion of an individual. Meanwhile, Dong et al. [71] proposed an Identity Consistency Transformer that learns simultaneously and identifies vectors for the inner face and another for the outer face; moreover, the model uses a novel consistency loss to drive both identities apart when their labels are different and to bring them closer when their labels are the same. Similarly, Nirkin et al. [72] identified deepfakes by looking for identity-to-identify inaccuracies between two identity vectors that represent the inner face region and its outer context. The identity vectors are obtained using two networks based on the Xception architecture and trained using a vanilla cross entropy loss. Focusing on temporal identity inconsistency, Liu et al. [73] introduced a model that captures the disparities of faces within video frames of the same person by encoding identity information in all frames to identity vectors and learning from these vectors the temporal embeddings, thus identifying inconsistencies. The proposed model integrates triplet loss for enhanced discrimination in learning temporal embeddings.

4.1.6. Detection Based on Facial Emotions

Despite the fact that deepfakes can produce convincing audio and video, it can be difficult to produce material that maintains coherence concerning high-level semantics, including emotions. Unnatural displays of emotion, as determined by characteristics like valence and arousal, where arousal indicates either heightened excitement or tranquility and valence represents positivity or negativity of the emotional state, can offer compelling proof that a video has been artificially created. Using the emotion inferred from the visage and vocalizations of the speaker, Hosler et al. [74] introduced an approach for identifying deepfakes. The suggested method makes use of long, short-term memory networks and visual descriptors to infer emotion from low-level audio emotion; a supervised classifier is then incorporated to categorize videos as real or fake using the predicted emotion. Leveraging the same high-level features, Conti et al. [75] focused on identifying deepfake speech tracks created using text-to-speech (TTS) algorithms that manipulate the emotional tone of the voice content. To extract emotional features, a Speech Emotion Recognition network trained on a speech dataset labeled with the speaker's emotional expression is employed, alongside a supervised classifier that receives emotional features as input and predicts the authenticity of the provided speech track as either genuine or deepfake.

4.1.7. Detection Based on Temporal Inconsistencies

To leverage temporal coherence to detect deepfakes, Zheng et al. [76] proposed an approach to reduce the spatial convolution kernel size to 1 while keeping the temporal convolution kernel size constant using a fully temporal convolution network in addition to a Transformer Network that explores the long-term temporal coherence. Pei et al. [77] exploited the temporal information in videos by incorporating a Bidirectional-LSTM model. Gu et al. [78] proposed a Region-Aware Temporal Filter module to generate temporal filters to distinct spatial areas by breaking down the dynamic temporal kernel into fundamental, region-independent filters. Additionally, region-specific aggregation weights are introduced to steer these regions in adaptively acquiring knowledge of temporal incongruities. The input video is split into multiple snippets to cover the long-term temporal dynamics. Inspired by how humans detect fake media through browsing and scrutinizing, Ru et al. [79] presented a model dubbed Bita-Net which consists of two pathways: one that checks the

temporal consistency by rapidly scanning the entire video, and a second pathway improved by an attention branch to analyze key frames of the video at a lower rate.

4.1.8. Detection Based on Spatial-Temporal Features

The forced mixing of the manipulated face in the generation process of deepfakes causes spatial distortions and temporal inconsistencies in crucial facial regions, which Sun et al. [80] proposed to reveal by extracting the displacement trajectory of the facial region. For the purpose of detecting fake trajectories, a fake trajectory detection network, utilizing a gated recurrent unit backbone in conjunction with a dual-stream spatial-temporal graph attention mechanism, is created. In order to detect the spatial-temporal abnormalities in the altered video trajectory, the network makes use of the extracted trajectory and explicitly integrates the important data from the input sequences. Lu et al. [81] proposed a detection method based on an improved Capsule Network and the fusion of temporal–spatial features. The optical flow algorithm effectively captures the temporal characteristics of manipulated videos, and the improved Capsule Network reaches a thorough conclusion by considering temporal–spatial features using weight initialization and updating on a dynamic routing algorithm. Meanwhile, Waseem et al. [82] described a dual-stream convolutional neural network strategy is employed, incorporating XceptionNet and 3DCNN, to capture spatial irregularities and temporal variations. Initially, MTCNN is employed for face detection and extraction from input video frames. Subsequently, 3DCNN and XceptionNet are utilized to extract features from facial images. Finally, fully connected layers and sigmoid layers determine the authenticity of the video.

4.2. Deep Learning Models for Deepfake Detection

Several advanced technologies have been employed in the domain of deepfake detection, such as machine learning [83–85] and media forensics-based approaches [86]. However, it is widely acknowledged that deep learning-based models currently exhibit the most remarkable performance in discerning between fabricated and authentic digital media. These models leverage sophisticated neural network architectures known as backbone networks, displayed in Figure 3, which have demonstrated exceptional efficacy in computer vision tasks. Prominent examples of such architectures include VGG [87], EfficientNet [88], Inception [89], CapsNet [90], and ViT [91], and are particularly renowned for their prowess in the feature extraction phase. Deep learning-based detection models go beyond conventional methods by incorporating additional techniques to further enhance their performance. One such approach is meta-learning, which enables the model to learn from previous experiences and adapt its detection capabilities accordingly. By leveraging meta-learning, these models become more proficient at recognizing patterns and distinguishing between genuine and manipulated content.

Furthermore, data augmentation plays a crucial role in training deep learning-based detection models. This technique involves augmenting the training dataset with synthetic or modified samples, which enhances the model's capacity to generalize and recognize diverse variations of deepfake media. Data augmentation enables the model to learn from a wider range of examples and improves its robustness against different types of manipulations. Attention mechanisms have also proven to be valuable additions to deep learning-based detection models. By directing the model's focus toward relevant features and regions of the input data, attention mechanisms enhance the model's discriminative power and improve its overall accuracy. These mechanisms help the model select critical details [92], making it more effective in distinguishing between real and fake media. Collectively, the combination of deep learning-based architectures, meta-learning, data augmentation, and attention mechanisms has significantly advanced the field of deepfake detection. These technologies work in harmony to equip models with the ability to identify and flag manipulated media with unprecedented accuracy.

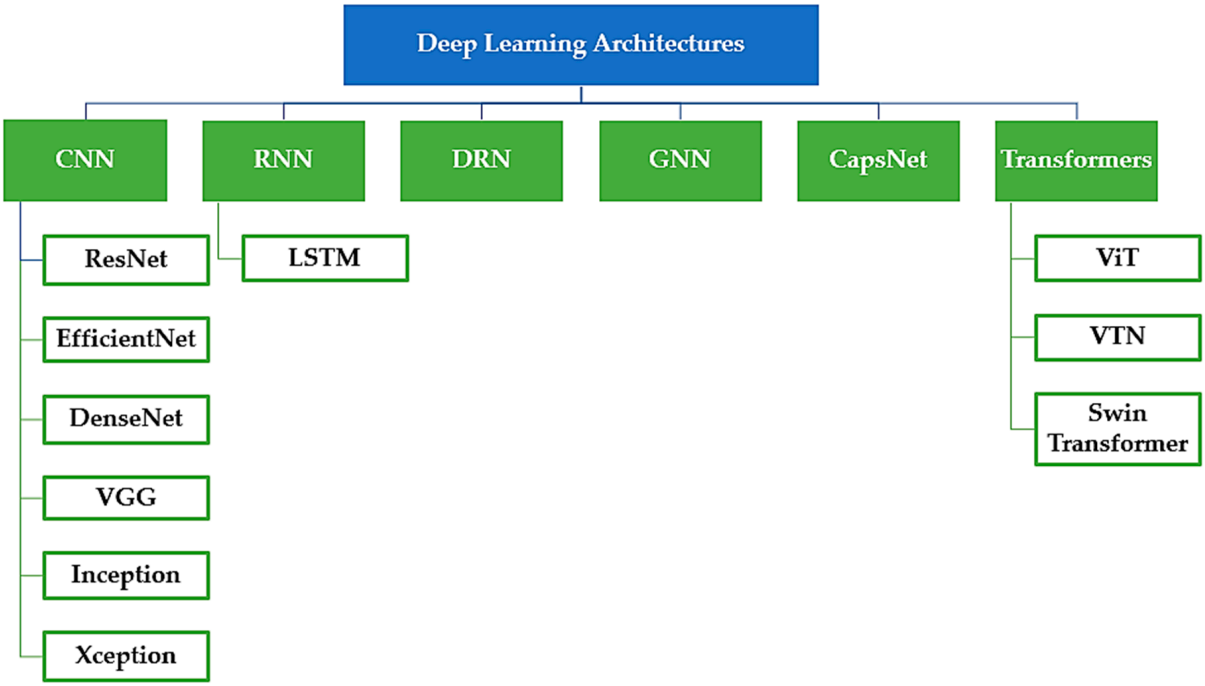

Figure 3. Overview of predominant deep learning architectures, networks, and frameworks employed in the development of deepfake detection models.

The Convolutional Neural Network is a powerful deep learning algorithm designed for image recognition and processing tasks. It consists of various levels, encompassing convolutional layers, pooling layers, and fully connected layers. There are different types of CNN models used in deepfake detection such as ResNet [93], short for Residual Network, which is an architecture that introduces skip connections to fix the vanishing gradient problem that occurs when the gradient diminishes significantly during backpropagation; these connections involve stacking identity mappings and skipping them, utilizing the layer's prior activations. This technique accelerates first training by reducing the number of layers in the network. The concept underlying this network is different from having the layers learn the fundamental mapping. Rather than directly defining the initial mapping as H(x), we let the network adapt and determine it, as shown in Figure 4.

$$F(x) := H(x) - x \text{ which gives } H(x) := F(x) + x.$$

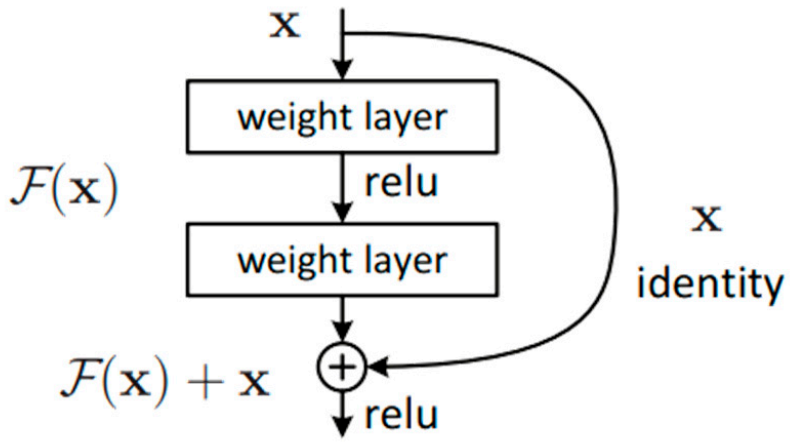

Figure 4. ResNet building block (source: [93]).

Another architecture based on CNNs is VGG [87], short for Visual Geometry Group, which comprises multiple layers. Instead of using large kernel sized filters, this architecture utilizes multiple filters with a kernel size of 3 × 3. The VGG16 architecture employs a doubling of filters at each convolutional layer, a fundamental design principle. However, a notable drawback of the VGG16 network is its substantial size, resulting in extended training times due to its depth and numerous fully connected layers. The model's file size exceeds 533 MB, rendering the implementation of a VGG network a time-intensive endeavor.

An additional significant CNN-based architecture in deepfake detection models is EfficientNet [88]. It has a scaling method that applies a uniform scaling approach to all dimensions of depth, width, and resolution. This is achieved by utilizing a compound coefficient. In Figure 5, the performance of EfficientNet is presented alongside other network architectures. The largest model within the EfficientNet series, EfficientNet B7, achieved remarkable results on both the ImageNet and CIFAR-100 datasets. Specifically, it achieved approximately 84.4% in top-1 accuracy and 97.3% in top-5 accuracy on the ImageNet dataset. Furthermore, this model was not only significantly more compact, being 8.4 times smaller, but also notably faster, with a speedup of 6.1 times compared to the prior leading CNN model. Additionally, it exhibited strong performance with 91.7% accuracy on the CIFAR-100 dataset and an impressive 98.8% accuracy on the Flowers dataset.

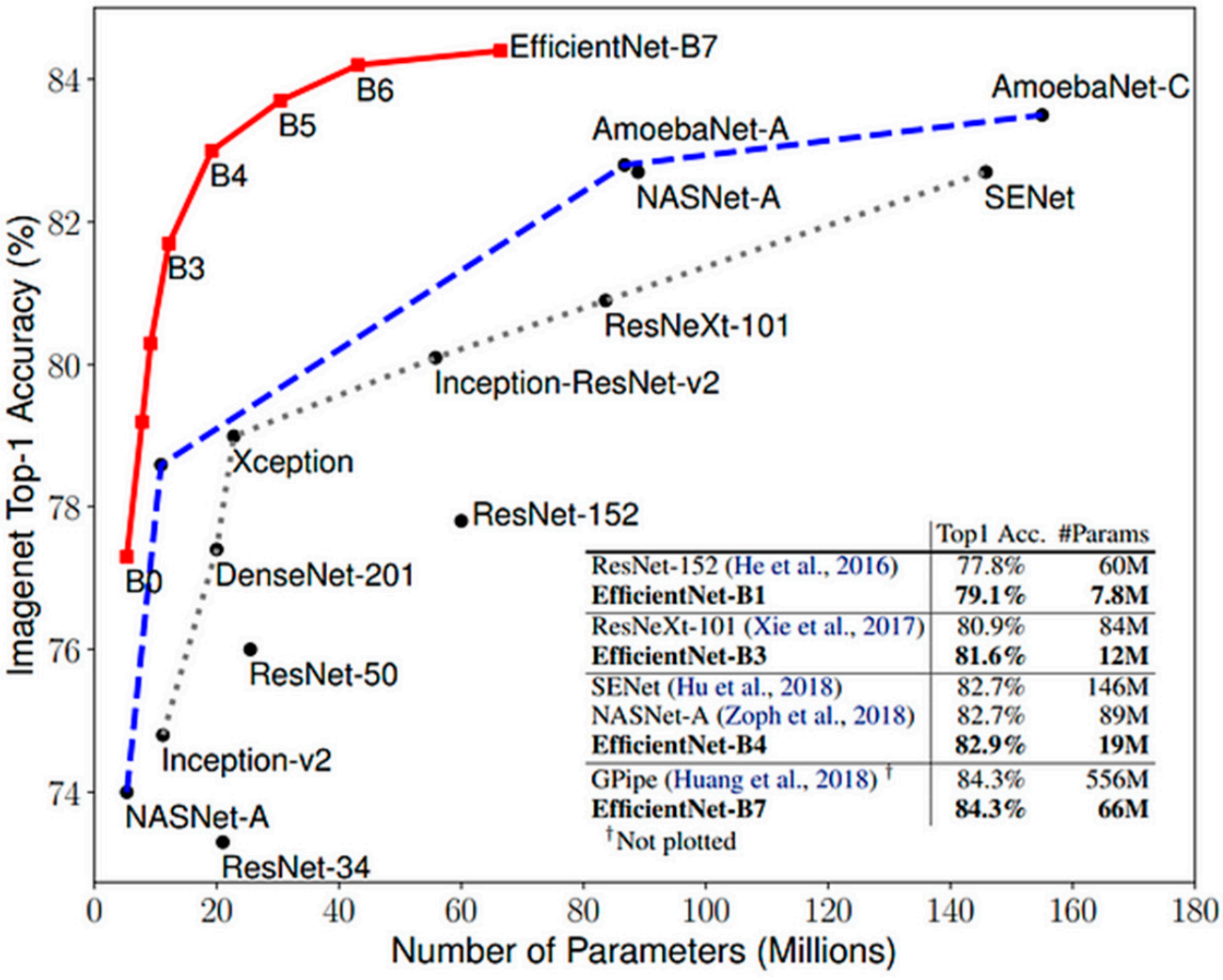

Figure 5. EfficientNet performance on the ImageNet dataset (source: [88]).

Inception [89] models help mitigate the computational cost and other overfitting in CNN architectures by utilizing stacked 1×1 convolutions for dimensionality reduction. Xception [94], developed by researchers at Google, is an advanced version of the Inception architecture. It offers a novel approach by reinterpreting Inception modules as an intermediate step between standard convolution and depthwise separable convolution. While the conventional convolution operation combines channel-wise and spatial-wise computations in a single step, depthwise separable convolution divides this process into two distinct steps. Firstly, it employs depthwise convolution to apply an individual convolutional filter to each input channel, and subsequently, pointwise convolution is employed to create a linear combination of the results obtained from the depthwise convolution.

An alternative to CNNs would be Capsule Networks [90] that are able to retrieve spatial information as well as other important details to avoid the information loss seen during pooling operations. Capsules exhibit equivariance characteristics and consist of a neural network that handles vectors as inputs and outputs, in contrast to the scalar values processed by CNNs. This unique attribute of capsules enables them to capture not only the features of an image, but also its deformations and various viewing conditions. Within a capsule network, each capsule comprises a cluster of neurons, with each neuron's output signifying a distinct attribute of the same feature. This structure offers the advantage of recognizing the entire entity by first identifying its constituent parts.

Recurrent Neural Networks are a kind of neural network that handles sequential data by feeding it in a sequential manner. They are specifically designed to tackle the challenge of time-series data, where the input is a sequence of data points. In an RNN, the input not only includes the current data point but also the previous ones. This creates a directed graph structure between the nodes, following the temporal sequence of the data. Additionally, each neuron in an RNN has its own internal memory, which retains information from the computations performed on the previous data points. LSTM, or Long Short-Term Memory, is a specific type of recurrent neural network that addresses the challenge of long-term dependencies in sequential data by allowing more accurate predictions based on recent information. While traditional RNNs struggle as the gap between relevant information increases, LSTM networks excel at retaining information over extended periods. This capability makes LSTM particularly effective for processing, predicting, and classifying time-series data.

A new model that has emerged as a strong alternative to convolutional neural networks is the vision transformer [91]. ViT models exhibit exceptional performance, surpassing the state-of-the-art CNNs by nearly four times in both computational efficiency and accuracy. Transformers, which are non-sequential deep learning models, play a significant role in vision transformers. They utilize the self-attention mechanism, assigning varying degrees of importance to different segments of the input data. The Swin Transformer [95] is a type of ViTs that exhibits versatility in modeling at different scales and maintains linear computational complexity concerning image size. This advantageous combination of features enables the Swin Transformer to be well suited for a wide array of vision tasks, encompassing image classification, object detection, and semantic segmentation, among others. Another variant of transformers is Video Transformers [96], which are efficient for evaluating videos on a large scale, ensuring optimal utilization of computational resources and reduced wall runtime. This capability enables full video processing during test time, making VTNs particularly well-suited for handling lengthy videos. Table 2 shows some of the recent detection techniques.

Table 2. Summary of recent deepfake detection models, employed techniques, feature sets, datasets, and intra-dataset performance results.

Author	Features	Technique	Intra-Dataset Performance (%)	Dataset
Zhao et al. [97]	Spatial temporal	Xception, Video Transformer	ACC (DF = 98.9 F2F = 96.1 FS = 97.5 NT = 92.1)	FF++(LQ)
			ACC (DF = 99.6 F2F = 99.6 FS = 100 NT = 96.8)	FF++(HQ)
			ACC = 99.8	Celeb-DF
			ACC = 92.1	DFDC
Yu et al. [98]	Spatial temporal	Global Inconsistency View, Multi-timescale Local Inconsistency View	ACC = 98.86 AUC = 99.89	FF++
			ACC = 98.78 AUC = 99.81	DFD
			ACC = 95.93 AUC = 98.96	DFDC
			ACC = 99.64 AUC = 99.78	Celeb-DF
			ACC = 98.94 AUC = 99.27	DFR1.0
Yang, Z. et al. [99]	Attentional features from facial regions	3D-CNN, TGCN, Spatial-temporal Attention, Masked Relation Learner	ACC = 91.81	FF++(LQ)
			ACC = 93.82	FF++(HQ)
			AUC = 99.96	Celeb-DF
			AUC = 99.11	DFDC
Yang, W. et al. [62]	Audio-Visual Features	Temporal-Spatial Encoder, Multi-Modal Joint-Decoder	ACC = 95.3 AUC = 97.6	DefakeAVMiT
			ACC = 83.7 AUC = 89.2	FakeAVCeleb
			ACC = 91.4 AUC = 94.8	DFDC
Shang et al. [100]	Spatial temporal	Temporal convolutional network, Spatial Relation Graph Convolution Units, Temporal Attention Convolution Units	ACC (DF = 99.29 F2F = 97.14 FS = 100 NT = 95.36)	FF++(HQ), Celeb-DF, DFDC
Rajalaxmi et al. [101]	Spatial inconsistencies	Inception-ResNet-V2	ACC = 98.37	DFDC
Korshunov et al. [102]	Spatial temporal	Xception	ACC = 100.00	Celeb-DF
			ACC = 99.14	FF++
			AUC = 99.93	DFR1.0
			AUC = 96.57	HifiFace
Patel et al. [103]	Temporal inconsistencies	Dense CNN	ACC = 97.2	CelebA, FFHQ, GDWCT, AttGAN, STARGAN, StyleGAN, StyleGAN2
Pang et al. [104]	Spatial temporal	Bipartite Group Sampling, Inconsistency Excitation, Longstanding Inconsistency Excitation,	ACC = 85.61 AUC = 91.23	WildDeepfake
			ACC = 97.76 AUC = 99.57	FF++(HQ)
			ACC = 91.60 AUC = 96.55	FF++(LQ)
			ACC = 97.35 AUC = 99.75	DFDC
Mehra et al. [105]	Spatial temporal	3D-Residual-in-Dense Net	ACC (DF = 98.57 F2F = 97.84 FS = 94.62 NT = 96.05)	FF++
			AUC= 92.93	Celeb-DF

Table 2. *Cont.*

Author	Features	Technique	Intra-Dataset Performance (%)	Dataset
Lu et al. [81]	Spatial temporal	VGG Capsule Networks	ACC = 94.07	Celeb-DF, FF++
Liu et al. [73]	Identity information	Encoder, RNN	AUC (FF++ = 99.95)	FF++, DFD, DFR1.0, Celeb-DF
Lin et al. [106]	Face semantic information	EfficientNet-b4 ViT	AUC = 99.80	Celeb-DF
			AUC = 88.47	DFDC
			ACC = 90.74 AUC = 94.86	FF++(LQ)
			ACC = 82.63	WildDeepfake
Liang et al. [53]	Facial geometry features	Facial geometry prior module, CNN-LSTM	ACC = 99.60	FF++
			ACC = 97.00	DFR1.0
			ACC = 82.84	Celeb-DF
			ACC = 94.68	DFD
Khalid et al. [107]	Spatial inconsistencies	Swin Y-Net Transformers	ACC (DF = 97.12 F2F = 95.73 FS = 92.10 NT = 79.90)	FF++
			AUC (DF = 97.00 F2F = 97.00 FS = 93.00 NT = 83.00)	
			ACC = 97.91 AUC = 98.00	Celeb-DF
Chen et al. [65]	Bi-granularity artifacts	ResNet-18decoder	Celeb-DF AUC = 99.80 FF++ AUC = 99.39	Celeb-DF, FF++ DFD, DFDC-P, UADFV, DFTIMIT, WildDeepfake
Agarwal et al. [70]	Identity information	Action Units	AUC = 97.00	World Leaders Dataset, Wav2Lip, FaceSwap YouTube
Cai et al. [61]	Audio-visual inconsistencies	3DCNN 2DCNN	ACC = 99.00	LAV-DF
			ACC = 84.60	DFDC
Zhuang et al. [108]	Spatial inconsistencies	Vision Transformer	FF++ AUC = 99.33	FF++, Celeb-DF, DFD, DFDC
Yan et al. [109]	Spatial temporal frequency features	GNN	AUC = 91.90 ACC = 89.70	FF++(LQ)
			AUC = 99.50 ACC = 97.80	F++(HQ)
Saealal et al. [110]	Spatial temporal	VGG11	AUC = 0.9446	OpenForensics
Xu et al. [111]	Spatial inconsistencies	Supervised contrastive model, Xception	ACC = 93.47	FF++
Xia et al. [112]	Image texture	MesoNet	ACC = 94.10 AUC = 97.40	FF++
			ACC = 94.90 AUC = 94.30	Celeb-DF
			AUC = 96.50	UADFV
			AUC = 84.30	DFD
Wu, N. et al. [113]	Semantic features	Multisemantic path neural network	ACC = 76.31	FF++(LQ)
			ACC = 94.21	F++(HQ)
			AUC = 99.52	TIMIT(LQ)
			AUC = 99.12	TIMIT(HQ)
Wu, H. et al. [114]	Spatial inconsistencies	Multistream Vision Transformer Network	ACC = 89.04	FF++(LQ)
			ACC = 99.31	FF++(HQ)

Table 2. *Cont.*

Author	Features	Technique	Intra-Dataset Performance (%)	Dataset
Waseem et al. [82]	Spatial temporal	XceptionNet and 3DCNN	FF++ ACC (DF = 95.55 F2F = 77.05 NT = 75.35)	FF++, DFTIMIT, DFD
Cozzolino et al. [115]	Audio-visual inconsistencies	ResNet-50	Avg AUC = 94.6	DFDC, DFTIMIT, FakeAVCeleb, KoDF
Wang, J. et al. [57]	Spatial-frequency domain	Multi-modal Multi-scale Transformers	ACC = 92.89 AUC = 95.31	FF++(LQ)
			ACC = 97.93 AUC = 99.51	FF++(HQ)
			AUC = 99.80	Celeb-DF
			AUC = 91.20	SR-DF
Wang, B. et al. [116]	Image grey space features	CNN Siamese network	ACC (DF = 84.14 F2F = 98.62 FS = 99.49 NT = 98.90)	FF++(LQ)
			ACC (DF = 95.79 F2F = 97.12 FS = 97.37 NT = 84.71)	FF++(HQ)
Saealal et al. [117]	Biological signals (Eye blinking)	Cascade CNN-LSTM-FCNs	ACC (DF = 94.65 F2F = 90.37 FS = 91.54 NT = 86.76)	FF++

Abbreviations FaceForensics++ (FF++), DeepFakes (DF), Face2Face2 (F2F), FaceSwap (FS), NeuralTextures (NT), DeeperForensics-1.0 (DFR1.0).

5. Datasets

In the context of deepfakes, datasets serve as the foundation for training, testing, and benchmarking deep learning models. The accessibility of reliable and diverse datasets plays a crucial role in the development and evaluation of deepfake techniques. A variety of important datasets, summarized in Table 3, have been curated specifically for deepfake research, each addressing different aspects of the problem and contributing to the advancement of the field. Figure 6 shows some of the widely used datasets in deepfake detection models' improvement.

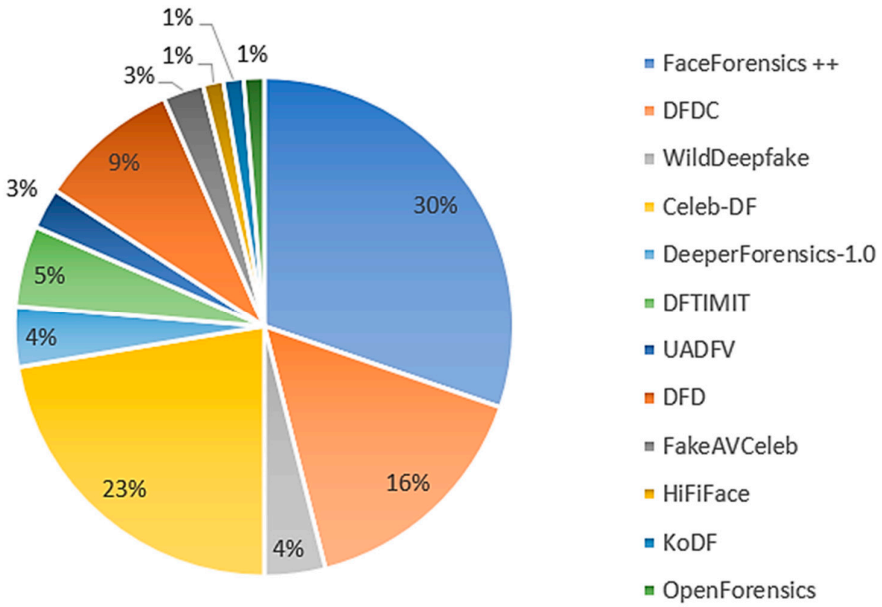

Figure 6. Frequency of usage of different deepfake datasets in the discussed detection models within this survey.

Table 3. Key characteristics of the most prominent and publicly available deepfake datasets.

Dataset	Year Released	Real Content	Fake Content	Generation Method	Modality
FaceForensics ++ [118]	2019	1000	4000	DeepFakes [119], Face2Face2 [37], FaceSwap [120], NeuralTextures [121], FaceShifter [34]	Visual
Celeb-DF (v2) [122]	2020	590	5639	DeepFake [122]	Visual
DFDC [123]	2020	23,654	104,500	DFAE, MM/NN, FaceSwap [120], NTH [124], FSGAN [125]	Audio/Visual
DeeperForensics-1.0 [126]	2020	48,475	11,000	DF-VAE [126]	Visual
WildDeepfake [127]	2020	3805	3509	Curated online	Visual
OpenForensics [128]	2021	45,473	70,325	GAN based	Visual
KoDF [129]	2021	62,166	175,776	FaceSwap [120], DeepFaceLab [51], FSGAN [125], FOMM [130], ATFHP [131], Wav2Lip [132]	Visual
FakeAVCeleb [133]	2021	500	19,500	FaceSwap [120], FSGAN [125], SV2TTS [134], Wav2Lip [132]	Audio/Visual
DeepfakeTIMIT [135]	2018	640	320	GAN based	Audio/Visual
UADFV [136]	2018	49	49	DeepFakes [119]	Visual
DFD [137]	2019	360	3000	DeepFakes [119]	Visual
HiFiFace [138]	2021	-	1000	HifiFace [138]	Visual

FaceForensics++ [118] is a well-known dataset used for deepfake detection that was provided in 2019 as an addition to the FaceForensics dataset, which was made available in 2018 and only included videos with altered facial expressions. Four subsets of the FF++ dataset are available: FaceSwap, Deepfake, Face2Face, and NeuralTextures. It includes 3000 edited videos in addition to 1000 original videos that were pulled from the YouTube-8M dataset. The dataset can be used to test deepfake detection strategies on both compressed and uncompressed videos because it is supplied in two different quality levels. The FF++ dataset has limits when it comes to spotting lip-sync deepfakes, and some videos might have color discrepancies near the modified faces.

DFDC [123], the deepfake detection challenge dataset hosted by Facebook, stands as the most extensive collection of face swap videos available and openly accessible. It contains over 100,000 total clips sourced from 3426 paid actors from diverse backgrounds, including different genders, ages, and ethnic groups.

DeeperForensics-1.0 [126] is a significant dataset available for detecting deepfakes that contains 50,000 original clips and 10,000 forged ones. These manipulated videos were generated using a conditional autoencoder called DF-VAE. The dataset includes a broad range of actor appearances and is designed to represent real-world scenarios more accurately by including a blend of alterations and disturbances, including compression, blurriness, noise, and other visual anomalies.

WildDeepfake [127] is a dataset that is widely recognized as a difficult one for deepfake detection. It features both authentic and deepfake samples obtained from the internet, which distinguishes it from other available datasets. While previous datasets have only included synthesized facial images, this dataset includes a variety of body types. However, there remains a need for a more comprehensive dataset that can generate full-body deepfakes to improve the robustness of deepfake detection models.

Celeb-DF [122] dataset is a collection of authentic and synthesized deepfake videos that are visually similar in quality to those that are commonly shared online. This dataset represents a significant expansion of its first version, which contained only 795 deepfake videos. Celeb-DF comprises 590 unaltered videos sourced from YouTube, featuring individuals of varying ages, ethnicities, and genders, along with 5639 associated deepfake videos, all of which were created using readily accessible YouTube excerpts featuring 59 famous personalities from diverse backgrounds. The deepfake videos were generated using an advanced synthesis method, resulting in more realistic and convincing deepfakes.

Finding fake faces among numerous genuine faces in scenes taken in the nature is a significant difficulty. OpenForensics [128] dataset was specifically created with face-wise rich annotations for the detection and segmentation of face forgeries. The OpenForensics dataset has a lot of potential for study in generic human face detection and deepfake prevention because of its extensive annotations. A total of 334 K human faces are depicted among 115 K photos in version 1.0.0. This collection includes numerous individuals with different origins, ages, genders, stances, positions, and face occlusions.

FakeAVCeleb [133] is a multimodal deepfake detection dataset that includes deepfake videos and cloned deepfake audio. It features diverse celebrities in terms of ethnicity, age, and gender balance. The dataset was evaluated using 11 different deepfake detection methods, including unimodal, ensemble-based, and multimodal approaches. To create deepfake videos, 500 real videos were used as sources and generated around 20,000 deepfake videos using various techniques like face-swapping and facial reenactment.

DeepfakeTIMIT [135] is a dataset containing 620 videos where faces were swapped using GAN-based techniques. It was created by selecting 16 pairs of similar-looking individuals from the VidTIMIT database, with two quality levels for each pair (64×64 and 128×128). The original audio tracks were retained without any alterations.

UADFV [136] dataset includes 98 videos, totaling 32,752 frames, evenly split between 49 real videos and 49 fake ones. Each video features a single subject and lasts around 11 s. Among these videos, there are 49 original real videos, which were manipulated to generate 49 Deep Fake videos.

DFD [137] or DeepFakeDetection is a dataset created by Google and Jigsaw and encompasses a wide range of scenes that consist of more than 363 genuine sequences featuring 28 paid actors across 16 different scenes. Additionally, it includes over 3000 manipulated videos.

HiFiFace [138] is a dataset that contains 1000 fake videos from FaceForensics++, meticulously adhering to the source and target pair configurations defined in FF++. Additionally, it includes 10,000 frames extracted from FF++ videos, facilitating quantitative testing.

KoDF [129] is an extensive compilation of synthesized and authentic videos primarily centered around Korean subjects. Its primary objective is to support the advancement of deepfake detection methods. This dataset comprises 62,166 authentic videos and 175,776 fake videos, featuring 403 different subjects.

One of the challenges faced by researchers in the field of deepfakes is the lack of comprehensive and diverse datasets for deepfake detection. Existing datasets either have limited diversity, meaning they do not cover a wide range of scenarios and variations, or only focus on basic forgery detection without capturing the intricacies and subtleties of advanced deepfakes. To address this problem and push the boundaries of deepfake detection, researchers and technology companies have taken up the task of constructing several benchmarks. These benchmarks serve as standardized datasets that encompass a broad range of facial variations, lighting conditions, camera angles, and other relevant factors. By including diverse samples, these benchmarks enable researchers to develop and evaluate advanced algorithms and techniques for detecting and analyzing deepfakes more effectively. To mention a few, ForgeryNet [139] is an extremely large deepfake benchmark with consistent annotations in both image and video data for four distinct tasks: Image Forgery Classification, Spatial Forgery Localization, Video Forgery Classification,

and Temporal Forgery Localization. It consists of 2.9 million images, 221,247 videos and 15 manipulation methods.

For the predominant focus on a single modality and limited coverage of forgery methods, current datasets for deepfake detection are primarily constrained when it comes to audio-visual deepfakes. DefakeAVMiT [62] is a dataset includes an ample amount of deepfake visuals paired with corresponding audios and generated by various deepfake methods affecting either modality. Alternatively, LAV-DF [61] consists of content-driven manipulations to help with the detection of content altering fake segments in videos due to the lack of suitable datasets for this task. It is important to note that the availability and creation of datasets are ongoing processes, with new datasets being introduced and existing ones being expanded or refined over time. The continuous development of diverse and representative datasets is crucial to ensure the robustness and generalizability of deepfake detection algorithms, as well as to keep up with the evolving techniques employed by malicious actors.

6. Challenges and Future Directions

Although deepfake detection has improved significantly, there are still a number of problems with the current detection algorithms that need to be resolved. The most significant challenge would be real-time detection of deepfakes and the implementation of detection models in diverse sectors and across multiple platforms. A challenge difficult to surmount due to several complexities, such as the computational power needed to detect deepfakes in real-time considering the massive amount of data shared every second on the internet and the necessity that these detection models be effective and have almost no instances of false positives. To attain this objective, one can leverage advanced learning techniques, such as meta-learning and metric learning, employ efficient architectures like transformers, apply compression techniques such as quantization, and make strategic investments in robust software and hardware infrastructure foundations.

In addition, detection methods encounter challenges intrinsic to deep learning, including concerns about generalization and robustness. Deepfake content frequently circulates across social media platforms after undergoing significant alterations like compression and the addition of noise. Consequently, employing detection models in real-world scenarios might yield limited effectiveness. To address this problem, several approaches have been explored to strengthen the generalization and robustness of detection models, such as feature restoration, attention guided modules, adversarial learning and data augmentation. Additionally, when it comes to deepfakes, the lack of interpretability of deep learning models becomes particularly problematic, making it challenging to directly grasp how they arrive at their decisions. This lack of transparency can be concerning, especially in critical applications, such as forensics, where understanding the reasoning behind a model's output is important for accountability, trust, and safety. Furthermore, since private data access may be necessary, detection methods raise privacy issues.

The quality of the deepfake datasets is yet another prominent challenge in deepfake detection. The development of deepfake detection techniques is made possible by the availability of large-scale datasets of deepfakes. The content in the available datasets, however, has some noticeable visual differences from the deepfakes that are actually being shared online. Researchers and technology companies such as Google and Facebook constantly put forth datasets and benchmarks to improve the field of deepfake detection. A further threat faced by detection models is adversarial perturbations that can successfully deceive deepfake detectors. These perturbations are strategically designed to exploit vulnerabilities or weaknesses in the underlying algorithms used by deepfake detectors. By introducing subtle modifications to the visual or audio components of a deepfake, adversarial perturbations can effectively trick the detectors into misclassifying the manipulated media as real.

Deepfake detection algorithms, although crucial, cannot be considered the be-all end-all solution in the ongoing battle against the threat they pose. Recognizing this, numerous

approaches have emerged within the field of deepfakes that aim to not only identify these manipulated media but also provide effective means to mitigate and defend against them. These multifaceted approaches serve the purpose of not only detecting deepfakes but also hindering their creation and curbing their rapid dissemination across various platforms. One prominent avenue of exploration in combating deepfakes involves the incorporation of adversarial perturbations to obstruct the creation of deepfakes. An alternative method involves employing digital watermarking, which discreetly embeds data or signatures within digital content to safeguard its integrity and authenticity. Additionally, blockchain technology offers a similar solution by generating a digital signature for the content and storing it on the blockchain, enabling the verification of any alterations or manipulations to the content.

Moreover, increasing public knowledge of the existence and potential risks linked with deepfakes is essential. Education and media literacy initiatives can educate users on how to critically evaluate digital content, recognize signs of manipulation, and verify the authenticity of media before sharing or believing its content. By empowering individuals to be more discerning consumers of information, the impact of deepfakes can be mitigated. Lastly, governments and policymakers are working to develop regulations and laws that address the misuse of deepfake technology. These policies and legislative measures aim to prevent the creation and dissemination of malicious deepfakes, establish liability frameworks for their creation and distribution, and protect individuals' rights and privacy.

7. Conclusions

In conclusion, deepfake videos will get harder to detect as AI algorithms become more sophisticated. This survey paper has provided a comprehensive overview encompassing the realm of deepfake generation, the spectrum of deep learning architectures employed in detection, the latest advances in detection techniques, and the pivotal datasets tailored to advance this field of study, all in order to stay one step ahead in the race with generative artificial intelligence, curb the spread of false information, safeguard the integrity of digital content, and stop the damage that deepfakes can cause on a social, political, and economic level. The survey has also highlighted the importance of continued research and development in deepfake detection techniques. Despite the issues presented by deepfakes, this technology nevertheless shows potential for artistic uses in virtual communication, entertainment, and visual effects. Future work must continue to focus on finding a balance between utilizing deepfakes' beneficial potential and reducing their negative effects.

Author Contributions: Conceptualization, A.N., M.R., F.S. and N.K.; methodology, A.N.; formal analysis, A.N.; investigation, A.N.; writing—original draft preparation, A.N.; writing—review and editing M.R., N.K. and F.S.; supervision, M.R., F.S. and N.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

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

References

1. Hancock, J.T.; Bailenson, J.N. The Social Impact of Deepfakes. *Cyberpsychol. Behav. Soc. Netw.* **2021**, *24*, 149–152. [CrossRef] [PubMed]
2. Giansiracusa, N. *How Algorithms Create and Prevent Fake News: Exploring the Impacts of Social Media, Deepfakes, GPT-3, and More*; Apress: Berkeley, CA, USA, 2021; ISBN 978-1-4842-7154-4.
3. Fallis, D. The Epistemic Threat of Deepfakes. *Philos. Technol.* **2021**, *34*, 623–643. [CrossRef] [PubMed]
4. Karnouskos, S. Artificial Intelligence in Digital Media: The Era of Deepfakes. *IEEE Trans. Technol. Soc.* **2020**, *1*, 138–147. [CrossRef]
5. Ridouani, M.; Benazzouza, S.; Salahdine, F.; Hayar, A. A Novel Secure Cooperative Cognitive Radio Network Based on Chebyshev Map. *Digit. Signal Process.* **2022**, *126*, 103482. [CrossRef]

6. 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]
7. Seow, J.W.; Lim, M.K.; Phan, R.C.W.; Liu, J.K. A Comprehensive Overview of Deepfake: Generation, Detection, Datasets, and Opportunities. *Neurocomputing* **2022**, *513*, 351–371. [CrossRef]
8. Rana, M.S.; Nobi, M.N.; Murali, B.; Sung, A.H. Deepfake Detection: A Systematic Literature Review. *IEEE Access* **2022**, *10*, 25494–25513. [CrossRef]
9. Akhtar, Z. Deepfakes Generation and Detection: A Short Survey. *J. Imaging* **2023**, *9*, 18. [CrossRef]
10. Ahmed, S.R.; Sonuç, E.; Ahmed, M.R.; Duru, A.D. Analysis Survey on Deepfake Detection and Recognition with Convolutional Neural Networks. In Proceedings of the 2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), Ankara, Turkey, 9–11 June 2022; pp. 1–7.
11. 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]
12. Yu, P.; Xia, Z.; Fei, J.; Lu, Y. A Survey on Deepfake Video Detection. *IET Biom.* **2021**, *10*, 607–624. [CrossRef]
13. Mirsky, Y.; Lee, W. The Creation and Detection of Deepfakes: A Survey. *ACM Comput. Surv.* **2021**, *54*, 1–41. [CrossRef]
14. Masood, M.; Nawaz, M.; Malik, K.M.; Javed, A.; Irtaza, A. Deepfakes Generation and Detection: State-of-the-Art, Open Challenges, Countermeasures, and Way forward. *Appl. Intell.* **2023**, *53*, 3974–4026. [CrossRef]
15. Das, A.; Viji, K.S.A.; Sebastian, L. A Survey on Deepfake Video Detection Techniques Using Deep Learning. In Proceedings of the 2022 Second International Conference on Next Generation Intelligent Systems (ICNGIS), Kerala, India, 29–31 July 2022; pp. 1–4.
16. Lin, K.; Han, W.; Gu, Z.; Li, S. A Survey of DeepFakes Generation and Detection. In Proceedings of the 2021 IEEE Sixth International Conference on Data Science in Cyberspace (DSC), Shenzhen, China, 9–11 October 2021; pp. 474–478.
17. Chauhan, R.; Popli, R.; Kansal, I. A Comprehensive Review on Fake Images/Videos Detection Techniques. In Proceedings of the 2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 13–14 October 2022; pp. 1–6.
18. Khichi, M.; Kumar Yadav, R. A Threat of Deepfakes as a Weapon on Digital Platform and Their Detection Methods. In Proceedings of the 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), Khargpur, India, 6–8 July 2021; pp. 1–8.
19. Chaudhary, S.; Saifi, R.; Chauhan, N.; Agarwal, R. A Comparative Analysis of Deep Fake Techniques. In Proceedings of the 2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), Greater Noida, India, 17–18 December 2021; pp. 300–303.
20. Younus, M.A.; Hasan, T.M. Abbreviated View of Deepfake Videos Detection Techniques. In Proceedings of the 2020 6th International Engineering Conference "Sustainable Technology and Development" (IEC), Erbil, Iraq, 26–27 February 2020; pp. 115–120.
21. Sudhakar, K.N.; Shanthi, M.B. Deepfake: An Endanger to Cyber Security. In Proceedings of the 2023 International Conference on Sustainable Computing and Smart Systems (ICSCSS), Coimbatore, India, 10–12 July 2023; pp. 1542–1548.
22. Salman, S.; Shamsi, J.A.; Qureshi, R. Deep Fake Generation and Detection: Issues, Challenges, and Solutions. *IT Prof.* **2023**, *25*, 52–59. [CrossRef]
23. Khder, M.A.; Shorman, S.; Aldoseri, D.T.; Saeed, M.M. Artificial Intelligence into Multimedia Deepfakes Creation and Detection. In Proceedings of the 2023 International Conference on IT Innovation and Knowledge Discovery (ITIKD), Manama, Bahrain, 8–9 March 2023; pp. 1–5.
24. Kandari, M.; Tripathi, V.; Pant, B. A Comprehensive Review of Media Forensics and Deepfake Detection Technique. In Proceedings of the 2023 10th International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 15–17 March 2023; pp. 392–395.
25. Boutadjine, A.; Harrag, F.; Shaalan, K.; Karboua, S. A Comprehensive Study on Multimedia DeepFakes. In Proceedings of the 2023 International Conference on Advances in Electronics, Control and Communication Systems (ICAECCS), BLIDA, Algeria, 6–7 March 2023; pp. 1–6.
26. Mallet, J.; Dave, R.; Seliya, N.; Vanamala, M. Using Deep Learning to Detecting Deepfakes. In Proceedings of the 2022 9th International Conference on Soft Computing & Machine Intelligence (ISCMI), Toronto, ON, Canada, 26–27 November 2022; pp. 1–5.
27. Alanazi, F. Comparative Analysis of Deep Fake Detection Techniques. In Proceedings of the 2022 14th International Conference on Computational Intelligence and Communication Networks (CICN), Al-Khobar, Saudi Arabia, 4–6 December 2022; pp. 119–124.
28. Xinwei, L.; Jinlin, G.; Junnan, C. An Overview of Face Deep Forgery. In Proceedings of the 2021 International Conference on Computer Engineering and Application (ICCEA), Nanjing, China, 25–27 June 2021; pp. 366–370.
29. Weerawardana, M.; Fernando, T. Deepfakes Detection Methods: A Literature Survey. In Proceedings of the 2021 10th International Conference on Information and Automation for Sustainability (ICIAfS), Negambo, Sri Lanka, 11–13 August 2021; pp. 76–81.
30. Swathi, P.; Sk, S. DeepFake Creation and Detection:A Survey. In Proceedings of the 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, 2–4 September 2021; pp. 584–588.
31. Zhang, T.; Deng, L.; Zhang, L.; Dang, X. Deep Learning in Face Synthesis: A Survey on Deepfakes. In Proceedings of the 2020 IEEE 3rd International Conference on Computer and Communication Engineering Technology (CCET), Beijing, China, 14–16 August 2020; pp. 67–70.

32. Shi, Y.; Liu, X.; Wei, Y.; Wu, Z.; Zuo, W. Retrieval-Based Spatially Adaptive Normalization for Semantic Image Synthesis. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 11214–11223.
33. Liu, M.; Ding, Y.; Xia, M.; Liu, X.; Ding, E.; Zuo, W.; Wen, S. STGAN: A Unified Selective Transfer Network for Arbitrary Image Attribute Editing. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 3668–3677.
34. Li, L.; Bao, J.; Yang, H.; Chen, D.; Wen, F. FaceShifter: Towards High Fidelity And Occlusion Aware Face Swapping. *arXiv* **2020**, arXiv:1912.13457.
35. Robust and Real-Time Face Swapping Based on Face Segmentation and CANDIDE-3. Available online: https://www.springerprofessional.de/robust-and-real-time-face-swapping-based-on-face-segmentation-an/15986368 (accessed on 18 July 2023).
36. Ferrara, M.; Franco, A.; Maltoni, D. The Magic Passport. In Proceedings of the IEEE International Joint Conference on Biometrics, Clearwater, FL, USA, 29 September–2 October 2014; pp. 1–7.
37. Thies, J.; Zollhöfer, 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, Seattle, WA, USA, 13–19 June 2020.
38. Zhang, J.; Zeng, X.; Wang, M.; Pan, Y.; Liu, L.; Liu, Y.; Ding, Y.; Fan, C. FReeNet: Multi-Identity Face Reenactment. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; IEEE: Seattle, WA, USA, 2020; pp. 5325–5334.
39. Wang, Y.; Song, L.; Wu, W.; Qian, C.; He, R.; Loy, C.C. Talking Faces: Audio-to-Video Face Generation. In *Handbook of Digital Face Manipulation and Detection: From DeepFakes to Morphing Attacks*; Rathgeb, C., Tolosana, R., Vera-Rodriguez, R., Busch, C., Eds.; Advances in Computer Vision and Pattern Recognition; Springer International Publishing: Cham, Switzerland, 2022; pp. 163–188, ISBN 978-3-030-87664-7.
40. Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; et al. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 105–114.
41. He, Z.; Zuo, W.; Kan, M.; Shan, S.; Chen, X. AttGAN: Facial Attribute Editing by Only Changing What You Want. *IEEE Trans. Image Process.* **2019**, *28*, 5464–5478. [CrossRef] [PubMed]
42. 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, 15–19 June 2019.
43. Choi, Y.; Uh, Y.; Yoo, J.; Ha, J.-W. StarGAN v2: Diverse Image Synthesis for Multiple Domains. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; IEEE: Seattle, WA, USA, 2020; pp. 8185–8194.
44. Choi, Y.; Uh, Y.; Yoo, J.; Ha, J.-W. StarGAN v2: Diverse Image Synthesis for Multiple Domains. Available online: https://arxiv.org/abs/1912.01865v2 (accessed on 8 October 2023).
45. 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, Seattle, WA, USA, 13–19 June 2020.
46. Natsume, R.; Yatagawa, T.; Morishima, S. RSGAN: Face Swapping and Editing Using Face and Hair Representation in Latent Spaces. In Proceedings of the ACM SIGGRAPH 2018 Posters, Vancouver, BC, Canada, 12 August 2018; pp. 1–2.
47. Prajwal, K.R.; Mukhopadhyay, R.; Philip, J.; Jha, A.; Namboodiri, V.; Jawahar, C.V. Towards Automatic Face-to-Face Translation. In *Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 15 October 2019*; Association for Computing Machinery: New York, NY, USA, 2019; pp. 1428–1436.
48. FaceApp: Face Editor. Available online: https://www.faceapp.com/ (accessed on 5 October 2023).
49. Reface—AI Face Swap App & Video Face Swaps. Available online: https://reface.ai/ (accessed on 5 October 2023).
50. DeepBrain AI—Best AI Video Generator. Available online: https://www.deepbrain.io/ (accessed on 5 October 2023).
51. Perov, I.; Gao, D.; Chervoniy, N.; Liu, K.; Marangonda, S.; Umé, C.; Dpfks, M.; Facenheim, C.S.; RP, L.; Jiang, J.; et al. DeepFaceLab: Integrated, Flexible and Extensible Face-Swapping Framework. *arXiv* **2021**, arXiv:2005.05535.
52. Make Your Own Deepfakes [Online App]. Available online: https://deepfakesweb.com/ (accessed on 5 October 2023).
53. Liang, P.; Liu, G.; Xiong, Z.; Fan, H.; Zhu, H.; Zhang, X. A Facial Geometry Based Detection Model for Face Manipulation Using CNN-LSTM Architecture. *Inf. Sci.* **2023**, *633*, 370–383. [CrossRef]
54. Li, G.; Zhao, X.; Cao, Y. Forensic Symmetry for DeepFakes. *IEEE Trans. Inf. Forensics Secur.* **2023**, *18*, 1095–1110. [CrossRef]
55. Hu, J.; Liao, X.; Gao, D.; Tsutsui, S.; Qin, Z.; Shou, M.Z. DeepfakeMAE: Facial Part Consistency Aware Masked Autoencoder for Deepfake Video Detection. *arXiv* **2023**, arXiv:2303.01740. [CrossRef]
56. Yang, J.; Xiao, S.; Li, A.; Lu, W.; Gao, X.; Li, Y. MSTA-Net: Forgery Detection by Generating Manipulation Trace Based on Multi-Scale Self-Texture Attention. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 4854–4866. [CrossRef]
57. Wang, J.; Wu, Z.; Ouyang, W.; Han, X.; Chen, J.; Lim, S.-N.; Jiang, Y.-G. M2TR: Multi-Modal Multi-Scale Transformers for Deepfake Detection. In *Proceedings of the ICMR—International Conference on Multimedia Retrieval, Newark, NJ, USA, 27–30 June 2022*; Association for Computing Machinery, Inc.: New York, NY, USA, 2022; pp. 615–623.
58. Xiao, S.; Yang, J.; Lv, Z. Protecting the Trust and Credibility of Data by Tracking Forgery Trace Based on GANs. *Digit. Commun. Netw.* **2022**, *8*, 877–884. [CrossRef]

59. Li, Y.; Chang, M.-C.; Lyu, S. In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking. In Proceedings of the International Workshop on Information Forensics and Security, WIFS, Hong Kong, China, 11–13 December 2018; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2019.
60. Hernandez-Ortega, J.; Tolosana, R.; Fierrez, J.; Morales, A. DeepFakesON-Phys: Deepfakes Detection Based on Heart Rate Estimation. *arXiv* **2020**, arXiv:2010.00400.
61. 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: Anonymous Submission Paper ID 73. In Proceedings of the International Conference on Digital Image Computing: Techniques and Applications, DICTA, Sydney, Australia, 30 November–2 December 2022; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2022.
62. Yang, W.; Zhou, X.; Chen, Z.; Guo, B.; Ba, Z.; Xia, Z.; Cao, X.; Ren, K. AVoiD-DF: Audio-Visual Joint Learning for Detecting Deepfake. *IEEE Trans. Inf. Forensics Secur.* **2023**, *18*, 2015–2029. [CrossRef]
63. Ilyas, H.; Javed, A.; Malik, K.M. AVFakeNet: A Unified End-to-End Dense Swin Transformer Deep Learning Model for Audio–Visual Deepfakes Detection. *Appl. Soft Comput.* **2023**, *136*, 110124. [CrossRef]
64. Huang, Y.; Juefei-Xu, F.; Guo, Q.; Liu, Y.; Pu, G. FakeLocator: Robust Localization of GAN-Based Face Manipulations. *IEEE Trans. Inf. Forensics Secur.* **2022**, *17*, 2657–2672. [CrossRef]
65. Chen, H.; Li, Y.; Lin, D.; Li, B.; Wu, J. Watching the BiG Artifacts: Exposing DeepFake Videos via Bi-Granularity Artifacts. *Pattern Recogn.* **2023**, *135*, 109179. [CrossRef]
66. Guarnera, L.; Giudice, O.; Battiato, S. DeepFake Detection by Analyzing Convolutional Traces. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 14 June 2020; pp. 2841–2850.
67. Cho, W.; Choi, S.; Park, D.K.; Shin, I.; Choo, J. Image-to-Image Translation via Group-Wise Deep Whitening-and-Coloring Transformation. Available online: https://arxiv.org/abs/1812.09912v2 (accessed on 8 October 2023).
68. 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. Available online: https://arxiv.org/abs/1711.09020v3 (accessed on 8 October 2023).
69. Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; Aila, T. Analyzing and Improving the Image Quality of StyleGAN. Available online: https://arxiv.org/abs/1912.04958v2 (accessed on 8 October 2023).
70. Agarwal, S.; Hu, L.; Ng, E.; Darrell, T.; Li, H.; Rohrbach, A. Watch Those Words: Video Falsification Detection Using Word-Conditioned Facial Motion. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, WACV, Waikoloa, HI, USA, 2–7 January 2023; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2023; pp. 4699–4708.
71. Dong, X.; Bao, J.; Chen, D.; Zhang, T.; Zhang, W.; Yu, N.; Chen, D.; Wen, F.; Guo, B. Protecting Celebrities from DeepFake with Identity Consistency Transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; IEEE Computer Society: Washington, DC, USA, 2022; Volume 2022, pp. 9458–9468.
72. Nirkin, Y.; Wolf, L.; Keller, Y.; Hassner, T. DeepFake Detection Based on Discrepancies Between Faces and Their Context. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 6111–6121. [CrossRef]
73. 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, WACV, Waikoloa, HI, USA, 2–7 January 2023; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2023; pp. 4680–4689.
74. Hosler, B.; Salvi, D.; Murray, A.; Antonacci, F.; Bestagini, P.; Tubaro, S.; Stamm, M.C. Do Deepfakes Feel Emotions? A Semantic Approach to Detecting Deepfakes via Emotional Inconsistencies. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; IEEE Computer Society: Washington, DC, USA, 2021; pp. 1013–1022.
75. Conti, E.; Salvi, D.; Borrelli, C.; Hosler, B.; Bestagini, P.; Antonacci, F.; Sarti, A.; Stamm, M.C.; Tubaro, S. Deepfake Speech Detection through Emotion Recognition: A Semantic Approach. In Proceedings of the ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 22–27 May 2022; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2022; Volume 2022, pp. 8962–8966.
76. Zheng, Y.; Bao, J.; Chen, D.; Zeng, M.; Wen, F. Exploring Temporal Coherence for More General Video Face Forgery Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2021; pp. 15024–15034.
77. Pei, S.; Wang, Y.; Xiao, B.; Pei, S.; Xu, Y.; Gao, Y.; Zheng, J. A Bidirectional-LSTM Method Based on Temporal Features for Deep Fake Face Detection in Videos. In Proceedings of the 2nd International Conference on Information Technology and Intelligent Control (CITIC 2022), Kunming, China, 15–17 July 2022; Nikhath, K., Ed.; SPIE: Washington, DC, USA, 2022; Volume 12346.
78. Gu, Z.; Yao, T.; Chen, Y.; Yi, R.; Ding, S.; Ma, L. Region-Aware Temporal Inconsistency Learning for DeepFake Video Detection. In Proceedings of the 31th International Joint Conference on Artificial Intelligence, Vienna, Austria, 23–29 July 2022; De Raedt, L., De Raedt, L., Eds.; International Joint Conferences on Artificial Intelligence: Vienna, Austria, 2022; pp. 920–926.
79. Ru, Y.; Zhou, W.; Liu, Y.; Sun, J.; Li, Q. Bita-Net: Bi-Temporal Attention Network for Facial Video Forgery Detection. In Proceedings of the 2021 IEEE International Joint Conference on Biometrics, IJCB, Shenzhen, China, 4–7 August 2021; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2021.

80. Sun, Y.; Zhang, Z.; Echizen, I.; Nguyen, H.H.; Qiu, C.; Sun, L. Face Forgery Detection Based on Facial Region Displacement Trajectory Series. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops, WACV, Waikoloa, HI, USA, 3–7 January 2023; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2023; pp. 633–642.
81. Lu, T.; Bao, Y.; Li, L. Deepfake Video Detection Based on Improved CapsNet and Temporal–Spatial Features. *Comput. Mater. Contin.* **2023**, *75*, 715–740. [CrossRef]
82. Waseem, S.; Abu-Bakar, S.R.; Omar, Z.; Ahmed, B.A.; Baloch, S. A Multi-Color Spatio-Temporal Approach for Detecting DeepFake. In Proceedings of the 2022 12th International Conference on Pattern Recognition Systems, ICPRS, Saint-Etienne, France, 7–10 June 2022; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2022.
83. Matern, F.; Riess, C.; Stamminger, M. Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations. In Proceedings of the 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW), Waikoloa Village, HI, USA, 7–11 January 2019; pp. 83–92.
84. Ciftci, U.A.; Demir, I.; Yin, L. FakeCatcher: Detection of Synthetic Portrait Videos Using Biological Signals. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *9*, 1. [CrossRef]
85. Benazzouza, S.; Ridouani, M.; Salahdine, F.; Hayar, A. A Novel Prediction Model for Malicious Users Detection and Spectrum Sensing Based on Stacking and Deep Learning. *Sensors* **2022**, *22*, 6477. [CrossRef]
86. Verdoliva, L. Media Forensics and DeepFakes: An Overview. *IEEE J. Sel. Top. Signal Process.* **2020**, *14*, 910–932. [CrossRef]
87. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2015**, arXiv:1409.1556.
88. Tan, M.; Le, Q.V. *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks*; PMLR: Westminster, UK, 2020.
89. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going Deeper with Convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014.
90. Sabour, S.; Frosst, N.; Hinton, G.E. Dynamic Routing between Capsules. *arXiv* **2017**, arXiv:1710.09829.
91. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image Is Worth 16 $\times$ 16 Words: Transformers for Image Recognition at Scale. *arXiv* **2021**, arXiv:2010.11929.
92. Benazzouza, S.; Ridouani, M.; Salahdine, F.; Hayar, A. Chaotic Compressive Spectrum Sensing Based on Chebyshev Map for Cognitive Radio Networks. *Symmetry* **2021**, *13*, 429. [CrossRef]
93. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015.
94. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; IEEE: Honolulu, HI, USA, 2017; pp. 1800–1807.
95. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 17–11 October 2021.
96. Neimark, D.; Bar, O.; Zohar, M.; Asselmann, D. Video Transformer Network. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 17–11 October 2021.
97. Zhao, C.; Wang, C.; Hu, G.; Chen, H.; Liu, C.; Tang, J. ISTVT: Interpretable Spatial-Temporal Video Transformer for Deepfake Detection. *IEEE Trans. Inf. Forensics Secur.* **2023**, *18*, 1335–1348. [CrossRef]
98. Yu, Y.; Zhao, X.; Ni, R.; Yang, S.; Zhao, Y.; Kot, A.C. Augmented Multi-Scale Spatiotemporal Inconsistency Magnifier for Generalized DeepFake Detection. *IEEE Trans Multimed.* **2023**, *99*, 1–13. [CrossRef]
99. Yang, Z.; Liang, J.; Xu, Y.; Zhang, X.; He, R. Masked Relation Learning for DeepFake Detection. *IEEE Trans. Inf. Forensics Secur.* **2023**, *18*, 1696–1708. [CrossRef]
100. Shang, Z.; Xie, H.; Yu, L.; Zha, Z.; Zhang, Y. Constructing Spatio-Temporal Graphs for Face Forgery Detection. *ACM Trans. Web* **2023**, *17*, 1–25. [CrossRef]
101. Rajalaxmi, R.R.; Sudharsana, P.P.; Rithani, A.M.; Preethika, S.; Dhivakar, P.; Gothai, E. Deepfake Detection Using Inception-ResNet-V2 Network. In Proceedings of the 2023 7th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 23–25 February 2023; pp. 580–586.
102. Korshunov, P.; Jain, A.; Marcel, S. Custom Attribution Loss for Improving Generalization and Interpretability of Deepfake Detection. In Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 22–27 May 2022; IEEE: Singapore, 2022; pp. 8972–8976.
103. Patel, Y.; Tanwar, S.; Bhattacharya, P.; Gupta, R.; Alsuwian, T.M.; Davison, I.E.; Mazibuko, T.F. An Improved Dense CNN Architecture for Deepfake Image Detection. *IEEE Access* **2023**, *11*, 22081–22095. [CrossRef]
104. Pang, G.; Zhang, B.; Teng, Z.; Qi, Z.; Fan, J. MRE-Net: Multi-Rate Excitation Network for Deepfake Video Detection. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *33*, 3663–3676. [CrossRef]
105. Mehra, A.; Agarwal, A.; Vatsa, M.; Singh, R. Motion Magnified 3-D Residual-in-Dense Network for DeepFake Detection. *IEEE Trans. Biom. Behav. Iden. Sci.* **2023**, *5*, 39–52. [CrossRef]
106. Lin, H.; Huang, W.; Luo, W.; Lu, W. DeepFake Detection with Multi-Scale Convolution and Vision Transformer. *Digit. Signal Process. Rev. J.* **2023**, *134*, 103895. [CrossRef]

107. Khalid, F.; Akbar, M.H.; Gul, S. SWYNT: Swin Y-Net Transformers for Deepfake Detection. In Proceedings of the 2023 International Conference on Robotics and Automation in Industry (ICRAI), Peshawar, Pakistan, 3–5 March 2023; pp. 1–6.
108. Zhuang, W.; Chu, Q.; Tan, Z.; Liu, Q.; Yuan, H.; Miao, C.; Luo, Z.; Yu, N. UIA-ViT: Unsupervised Inconsistency-Aware Method Based on Vision Transformer for Face Forgery Detection. In *European Conference on Computer Vision*; Avidan, S., Brostow, G., Cisse, M., Farinella, G., Hassner, T., Eds.; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2022; Volume 13665, pp. 391–407.
109. Yan, Z.; Sun, P.; Lang, Y.; Du, S.; Zhang, S.; Wang, W. Landmark Enhanced Multimodal Graph Learning for Deepfake Video Detection. *arXiv* **2022**, arXiv:2209.05419. [CrossRef]
110. Saealal, M.S.; Ibrahim, M.Z.; Shapiai, M.I.; Fadilah, N. In-the-Wild Deepfake Detection Using Adaptable CNN Models with Visual Class Activation Mapping for Improved Accuracy. In Proceedings of the 2023 5th International Conference on Computer Communication and the Internet (ICCCI), Fujisawa, Japan, 23–25 June 2023; IEEE: Fujisawa, Japan, 2023; pp. 9–14.
111. Xu, Y.; Raja, K.; Pedersen, M. Supervised Contrastive Learning for Generalizable and Explainable DeepFakes Detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops, WACVW, Waikoloa, HI, USA, 4–8 January 2022; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2022; pp. 379–389.
112. Xia, Z.; Qiao, T.; Xu, M.; Wu, X.; Han, L.; Chen, Y. Deepfake Video Detection Based on MesoNet with Preprocessing Module. *Symmetry* **2022**, *14*, 939. [CrossRef]
113. Wu, N.; Jin, X.; Jiang, Q.; Wang, P.; Zhang, Y.; Yao, S.; Zhou, W. Multisemantic Path Neural Network for Deepfake Detection. *Secur. Commun. Netw.* **2022**, *2022*, 4976848. [CrossRef]
114. Wu, H.; Wang, P.; Wang, X.; Xiang, J.; Gong, R. GGViT:Multistream Vision Transformer Network in Face2Face Facial Reenactment Detection. In Proceedings of the 2022 26th International Conference on Pattern Recognition, Montreal, QC, Canada, 21–25 August 2022; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2022; Volume 2022, pp. 2335–2341.
115. 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, 17–24 June 2023; pp. 943–952. [CrossRef]
116. Wang, B.; Li, Y.; Wu, X.; Ma, Y.; Song, Z.; Wu, M. Face Forgery Detection Based on the Improved Siamese Network. *Secur. Commun. Netw.* **2022**, *2022*, 5169873. [CrossRef]
117. Saealal, M.S.; Ibrahim, M.Z.; Mulvaney, D.J.; Shapiai, M.I.; Fadilah, N. Using Cascade CNN-LSTM-FCNs to Identify AIaltered Video Based on Eye State Sequence. *PLoS ONE* **2022**, *17*, e0278989. [CrossRef]
118. Rössler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; Niessner, M. FaceForensics++: Learning to Detect Manipulated Facial Images. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 November 2019; pp. 1–11.
119. GitHub—Deepfakes/Faceswap: Deepfakes Software for All. Available online: https://github.com/deepfakes/faceswap (accessed on 10 October 2023).
120. GitHub—MarekKowalski/FaceSwap: 3D Face Swapping Implemented in Python. Available online: https://github.com/MarekKowalski/FaceSwap/ (accessed on 10 October 2023).
121. Thies, J.; Zollhöfer, M.; Nießner, M. Deferred Neural Rendering: Image Synthesis Using Neural Textures. Available online: https://arxiv.org/abs/1904.12356v1 (accessed on 10 October 2023).
122. Li, Y.; Yang, X.; Sun, P.; Qi, H.; Lyu, S. Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 3204–3213.
123. 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.
124. GitHub—Cuihaoleo/Kaggle-Dfdc: 2nd Place Solution for Kaggle Deepfake Detection Challenge. Available online: https://github.com/cuihaoleo/kaggle-dfdc (accessed on 10 October 2023).
125. 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.
126. 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, Seattle, WA, USA, 13–19 June 2020.
127. 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. [CrossRef]
128. 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 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, BC, Canada, 11–17 October 2021; pp. 10097–10107.
129. Kwon, P.; You, J.; Nam, G.; Park, S.; Chae, G. KoDF: A Large-Scale Korean DeepFake Detection Dataset. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, BC, Canada, 11–17 October 2021; pp. 10724–10733.
130. 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.

131. Yi, R.; Ye, Z.; Zhang, J.; Bao, H.; Liu, Y.-J. Audio-Driven Talking Face Video Generation with Learning-Based Personalized Head Pose. *arXiv* **2020**, arXiv:2002.10137.
132. Prajwal, K.R.; Mukhopadhyay, R.; Namboodiri, V.; 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 October 2020; pp. 484–492.
133. Khalid, H.; Tariq, S.; Woo, S.S. FakeAVCeleb: A Novel Audio-Video Multimodal Deepfake Dataset. *arXiv* **2021**, arXiv:2108.05080.
134. Jia, Y.; Zhang, Y.; Weiss, R.J.; Wang, Q.; Shen, J.; Ren, F.; Chen, Z.; Nguyen, P.; Pang, R.; Moreno, I.L.; et al. Transfer Learning from Speaker Verification to Multispeaker Text-to-Speech Synthesis. Available online: https://arxiv.org/abs/1806.04558v4 (accessed on 10 October 2023).
135. Korshunov, P.; Marcel, S. DeepFakes: A New Threat to Face Recognition? Assessment and Detection. *arXiv* **2018**, arXiv:1812.08685.
136. 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, 12–17 May 2019; pp. 8261–8265.
137. Contributing Data to Deepfake Detection Research—Google Research Blog. Available online: https://blog.research.google/2019/09/contributing-data-to-deepfake-detection.html (accessed on 5 October 2023).
138. 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.
139. 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, 21–25 June 2021; IEEE Computer Society: Washington, DC, USA, 2021; pp. 4358–4367.