electronics

MDPI

Review

A Comprehensive Review of DeepFake Detection Using Advanced Machine Learning and Fusion Methods

Gourav Gupta [1,†], Kiran Raja [1], Manish Gupta [2,*,†], Tony Jan [2], Scott Thompson Whiteside [2] and Mukesh Prasad [3]

[1] Department of Computer Science, Norwegian University of Science and Technology (NTNU), Teknologivegen 22, 2815 Gjøvik, Norway
[2] Artificial Intelligence and Optimization Research Centre, Design and Creative Technology, Torrens University, 46 Mountain Street, Ultimo, NSW 2007, Australia
[3] Faculty of Engineering and IT (FEIT), University of Technology Sydney (UTS), 15 Broadway, Ultimo, NSW 2000, Australia
* Correspondence: manish.gupta@torrens.edu.au
† These authors contributed equally to this work.

Abstract: Recent advances in Generative Artificial Intelligence (AI) have increased the possibility of generating hyper-realistic DeepFake videos or images to cause serious harm to vulnerable children, individuals, and society at large with misinformation. To overcome this serious problem, many researchers have attempted to detect DeepFakes using advanced machine learning techniques and advanced fusion techniques. This paper presents a detailed review of past and present DeepFake detection methods with a particular focus on media-modality fusion and machine learning. This paper also provides detailed information on available benchmark datasets in DeepFake detection research. This review paper addressed the 67 primary papers that were published between 2015 and 2023 in DeepFake detection, including 55 research papers in image and video DeepFake detection methodologies and 15 research papers on identifying and verifying speaker authentication. This paper offers lucrative information on DeepFake detection research and offers a unique review analysis of advanced machine learning and modality fusion that sets it apart from other review papers. This paper further offers informed guidelines for future work in DeepFake detection utilizing advanced state-of-the-art machine learning and information fusion models that should support further advancement in DeepFake detection for a sustainable and safer digital future.

Keywords: DeepFake detection; advanced machine learning in DeepFake detection; modality fusion in DeepFake detection; comprehensive review of DeepFake detection

Citation: Gupta, G.; Raja, K.; Gupta, M.; Jan, T.; Whiteside, S.T.; Prasad, M. A Comprehensive Review of DeepFake Detection Using Advanced Machine Learning and Fusion Methods. *Electronics* **2024**, *13*, 95. https://doi.org/10.3390/electronics13010095

Academic Editors: Enjie Liu and Hongqing Yu

Received: 26 October 2023
Revised: 15 December 2023
Accepted: 21 December 2023
Published: 25 December 2023

1. Introduction

DeepFakes are causing significant concern among the general public. For instance, fake videos created by fraudsters can easily deceive the general public [1]. Such fake videos can spread virally on social media, causing irreversible harm to targeted individuals or organizations (e.g., high-profile personalities or a company with significant brand value). A more sinister threat emerges when DeepFakes are used to create child pornography or sexually explicit fake content [2,3].

Generally, humans cannot distinguish a real video from a DeepFake with the naked eye (or ears) [4]. On a superficial level, DeepFakes are created by combining several techniques, such as merging, combining, replacing, and superimposing images and video clips to create fake videos [5], making them appear real. Taking advantage of more recent AI techniques such as generative adversarial networks (GANs), DeepFakes can now generate hyper-realistic content by incorporating audio into the video, thereby not only altering the visual content but also making it realistic in terms of audio [6].

Several approaches have recently been proposed to detect such manipulated content by analyzing spatial and frequency information in images, as well as temporal and frequency information from audio and video. To advance the state of the art in detecting DeepFakes, several benchmarking datasets have been made available to the public. By leveraging these databases and existing approaches, state-of-the-art methods have evolved to exploit the concept of information fusion, providing robust detection of fake media.

Although there are numerous surveys on DeepFake detection (DFD) [7] detailing advances, challenges, and potential future work, this paper focuses on a sub-topic of media modality fusion in DFD, thus complementing existing review works. In this paper, we delve into existing approaches in DFD, listing relevant works and benchmarking databases and their respective results. Furthermore, we discuss challenges and potential future work aimed at pushing the boundaries of the current state of the art in DeepFake detection.

In the remainder of this paper, Section 2 provides a detailed overview of the available datasets in DFD, while Section 3 explores a number of relevant works using fusion for DFD in videos. This is followed by Section 4, which focuses on audio. Finally, Sections 5 and 6 delve into more advanced topics in DFD, concluding with remarks and directions for future research work.

2. Benchmark Datasets

In this digital era, "Digital Transformation" has become a new epicenter across the globe for its diverse lucrative applications, such as facial attendance, centralized data lake, and intelligent robotic automation. It provides ease and intelligence for applications that connect the daily activities of human beings by leveraging advanced technologies. Since 2018, rapid growth in modern generative models has been observed because of the synthesis of vision and visual-related domains such as face synthesis, frame synthesis, and tone synthesis. Considering the harmful effects of manipulated images and videos on individuals and society, several multinational companies (MNCs) and universities have generated their own synthesized "DeepFake Dataset" in order to implement deep-learning-based advanced pipelines for detecting fake videos or images.

Regularly updating benchmark datasets with diverse and evolving DeepFake content ensures that detection models are tested against a representative range of deceptive techniques. Each benchmark dataset or pertaining paper contains the evaluation metrics to determine the reliability of that dataset for further comparison with the improved version of DeepFake's algorithm. Although each dataset has itself gone through the training and testing phases, benchmarking is to be performed to show the outperformance of the existing or new detection algorithm on the updated dataset.

The first DeepFake dataset, named as "Face Forensic DeepFake (FF-DF)" was released in 2018 and it was added to the repository to improve detection accuracy. Table 1 shows the chronological progress in the arrival of the DeepFake Dataset reported by several researchers. Table 1 presents a list of various DeepFake datasets reported by several researchers.

Table 1. DeepFake datasets (DFDs) generated since 2018.

Dataset	Pristine Video	Fake Video	Total Videos	Release Date
FF-DF	1000	1000	2000	18-March
UADFV	49	49	98	18-November
DF-TIMIT	320	620	940	18-December
FF++ DF	1000	1000	2000	19-August
Google DFD	3000	3000	6000	19-September
DFDC-Preview Dateset	1131	4113	5214	19-September
Celeb-DF	590	5639	6229	19-November
DeeperForensics-1.0	50,000	10,000	60,000	20-June
DFDC-Full Dataset	23,654	104,500	128,154	20-June

The total number of DeepFake videos was 3038 (1669 fake videos + 1369 pristine videos) in 2018, but the total number of videos was 188,154 (114,500 fake videos + 73,654 pristine videos) in 2020. It is clear that the Deepfake Detection Challenge (DFDC)-Full Dataset had the largest corpus of the DeepFake datasets and the UADFV dataset had the least. The size of the DFD benchmark datasets further increased to over 100,000 videos in each dataset including DF-Platter in 2023 [8,9]. Most research utilized both past and present benchmark datasets in evaluating their DFD performance for fair comparisons.

3. Video and Image Modality Fusion in DeepFake Detection

Generally, DeepFake detection can be classified into two categories: temporal and spatial analysis for video DFD and frame forgery analysis for image DFD.

Li et al. [10] described a novel approach for detecting fake or manipulated faces in pristine images or videos. They observed one of the vital features of human facial activity, that is, eye blinking rate, for authenticating the physiological signal that is not properly incorporated in the synthesized fake videos, as demonstrated in Figure 1.

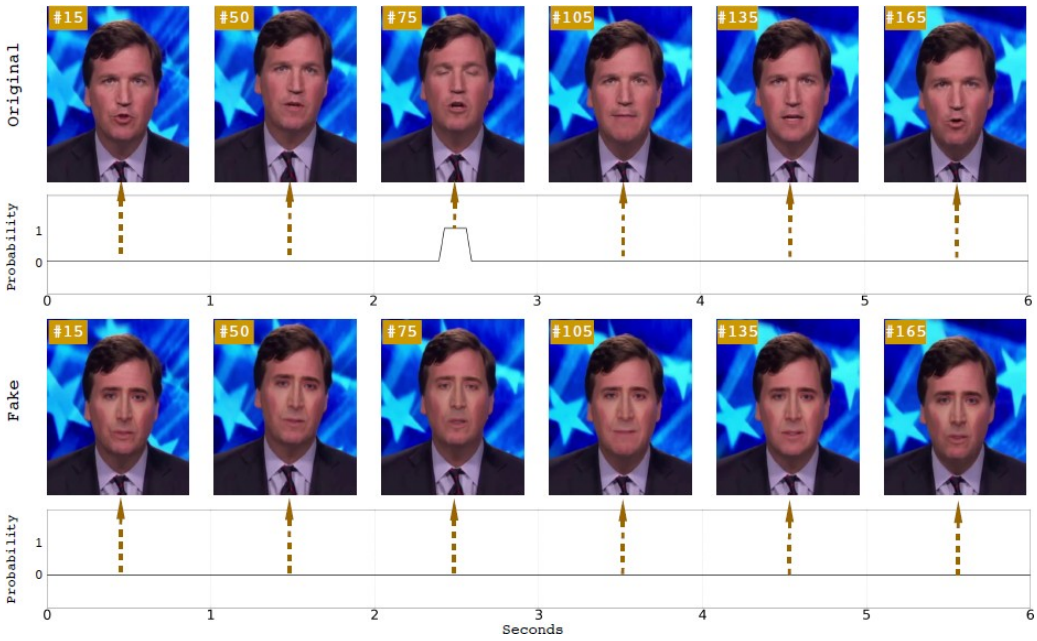

Figure 1. Eye-blinking detection on an original video (**Top**) and a DeepFake-generated video (**Bottom**).

In this study, the blinking rates of the eyes in pristine video were studied and compared with those of DeepFake videos. Using this novel method, the final outcome demonstrated that the proposed DeepFake detection mechanism can identify a synthesized or fake video when the blinking rate of the eyes is observed to be abnormal or anomalous. Figure 1 depicts the frame-by-frame observation of eye blinking of a human being in an original video and a DeepFake video, where the authors calculated the average time duration between consecutive eye blinks and the average time of eye blinking to be noticed for detecting real or fake videos. This method incorporates two stages: (a) face detection using image or frame, facial landmarks, face alignment, and eye region extraction, and (b) feeding features of the first stage into a long-term recurrent convolutional network (LRCN) to count the number of the eye blinks, as shown in Figure 2.

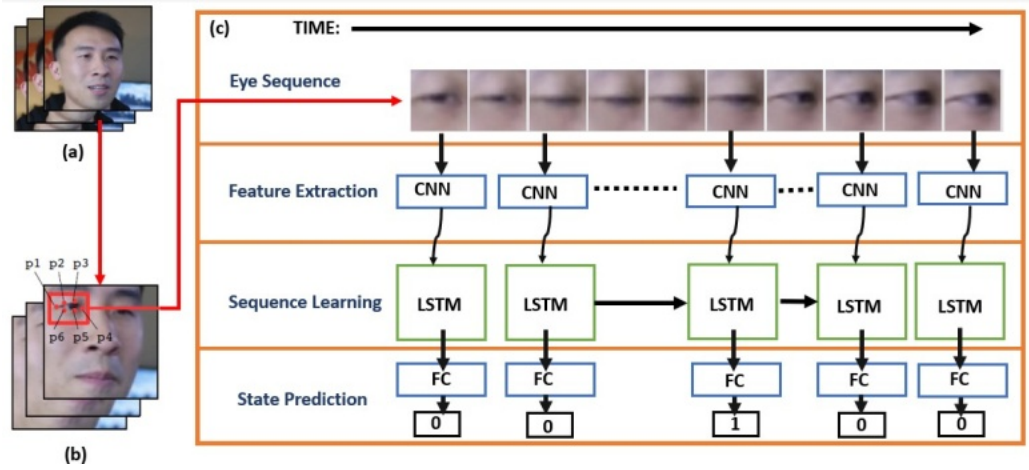

Figure 2. Overview of LRCN method. (**a**) is the original sequence. (**b**) is the sequence after face alignment and passed to (**c**) LRCN, which consists of three parts: feature extraction, sequence learning, and state prediction.

Afchar et al. [11] focused on analyzing the mesoscopic properties of images using detection systems based on a deep learning approach. They integrated two different activation functions and implemented two detection methods, Meso-4 and MesoInception-4, to distinguish between fake and real videos or images.

In Meso-4, they employed four successive layers of convolution and pooling, followed by a dense network with one hidden layer that utilized the Rectified Linear Unit (ReLU) activation function to improve generalization, as illustrated in Figure 3. However, in the MesoInception-4 architecture, the authors replaced the initial two layers of convolution with inception models and then applied them to the DeepFake and Face2Face datasets for evaluation, as depicted in Figure 4. The result showed a very high success rate in detection i.e., 98% for the DeepFake Dataset and 95% for the Face2Face dataset.

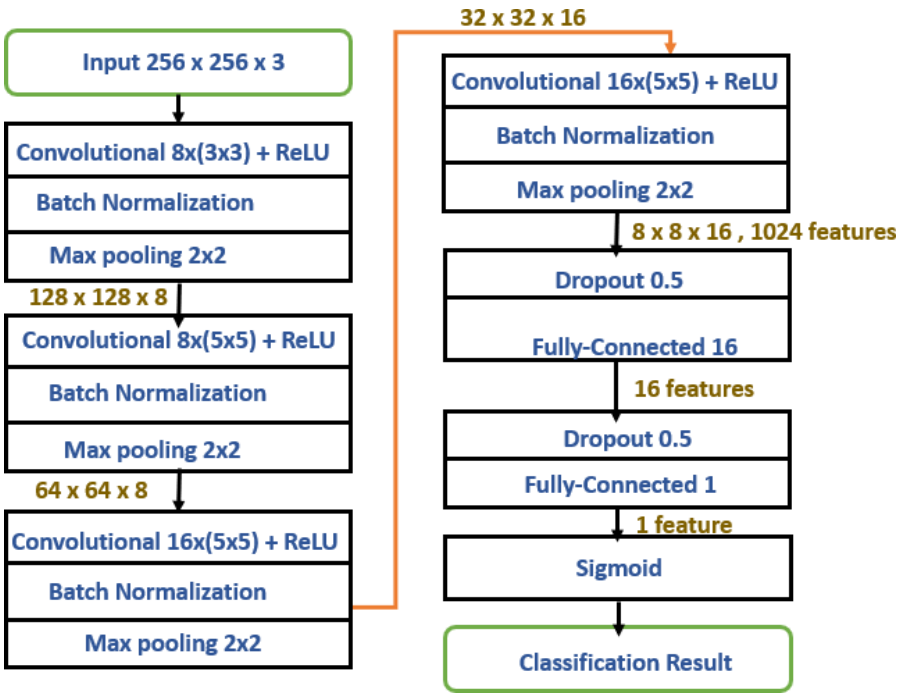

Figure 3. The network architecture of Meso-4.

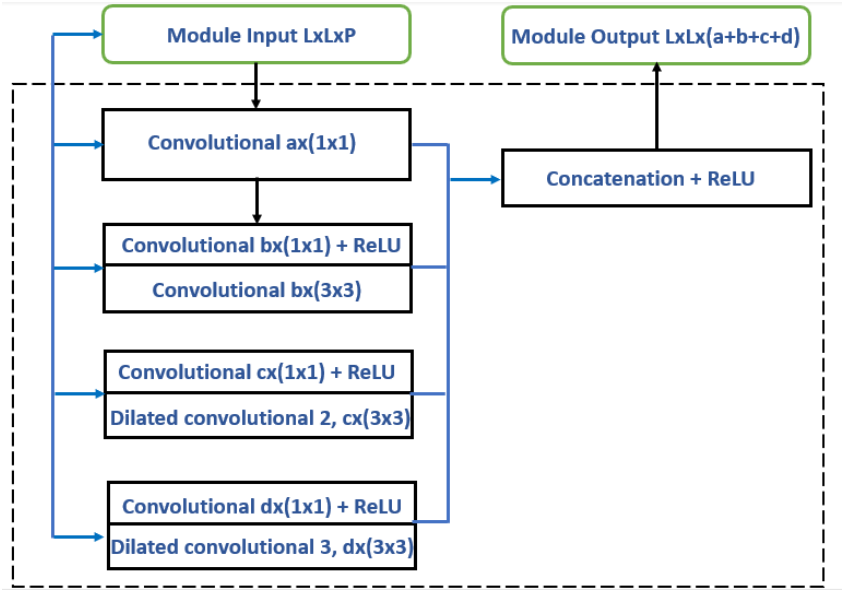

Figure 4. The network architecture of MesoInception-4.

Hinton et al. [12] addressed some major limitations of convolutional neural networks (CNNs) and proposed a foundation for a novel capsule architecture. Nguyen et al. [13] adopted the idea of capsule architecture and extended their work to detect different kinds of forgery in images and videos in addition to replay attacks. The authors [13] incorporated the state-of-the-art Deep Convolutional Neural Networks (DCNNs) and compared the outcomes of this method with other benchmarking methods. This method is widely used with dynamic routing algorithms and expectation maximization routing algorithms.

In this approach, the video stream is split into frames in the pre-processing phase, and then the detection, extraction, and scaling of faces from images are executed as input to the next phase. In the second phase, the extracted faces were fed to the VGG-19 network to extract latent features, which were later used as inputs to the Capsule Network. In the third phase, the Capsule Network, as shown in Figure 5, is executed to detect a forgery in images and videos, and post-processing is used to calculate the average probabilities for generating the final result. The Capsule-Forensics-Noise was 95.93% for the DeepFake Dataset at the frame level and 99.23% for the DeepFake Dataset at the video level.

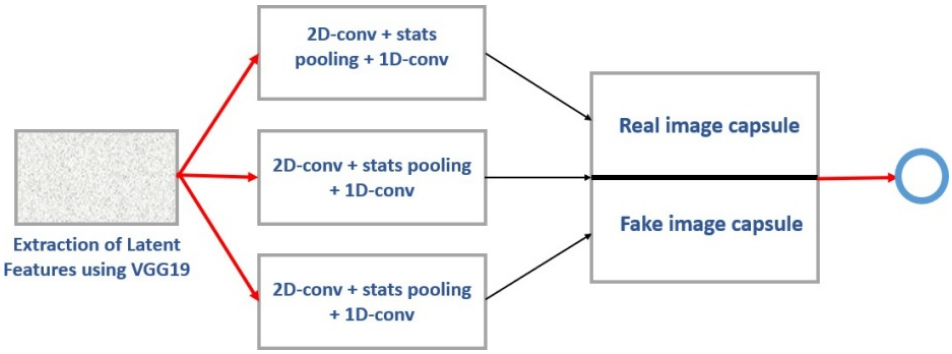

Figure 5. The architecture of Capsule Network.

Rossler et al. [14] proposed an automated pipeline to detect fake faces from images or videos. In this method, a tracking algorithm was used to trace and track the human face of videos or images, and then, fed it to different classifiers for detecting the forgery whether it persists in the videos or not. The authors [14] selected four DeepFake datasets, including DeepFakes, Face2Face, FaceSwap, and NeuralTextures, along with a pristine dataset to evaluate precision.

Detecting DeepFakes using divergent classifiers, such as Steganalysis Features and Support Vector Machine (SVM), Cozzolino et al., Bayar and Stamm, and Rahmouni utilized MesoNet, XceptionNet, and XceptionNet full images. While applying these classifiers randomly to a divergent range of video qualities, they found that the XceptionNet classifier outperformed other classifiers or combinations. The binary precision values on low-quality trained XceptionNet is 96.36% for DeepFakes (DFs), 86.86% for Face2Face (F2F), 90.29% for FaceSwap (FS), 52.4% for NeuralTextures (NT), and 52.04% for pristine images (Real-set).

Dolhansky et al. [15] implemented three detection models using different features on the DeepFake Detection Challenge (DFDC) dataset that consists of 5k videos, including original and fake clips. The sample dataset of DFDC is shown in Figure 6. In the first method, they applied a light-weighted DNN model, that is, TamperNet, which consists of six convolutional layers and one fully connected layer on the DeepFake Dataset, which is used to detect acute level manipulations on images such as cut-and-pasted objects; this method also identifies forgery in digitally fabricated images, including face swaps.

Figure 6. Examples of face swaps.

In the second method, they implemented two other detection models using XceptionNet on a face dataset and a full-image dataset on forensic data. In these frame-based models, there are two thresholds applied to the sampled frame per second of video: (1) a per-frame detection threshold and (2) a threshold that specifies how many frames must exceed the per-frame threshold to identify a video as fake. They evaluated the frame-per-video threshold over frames with a detectable face. During the validation, it was clearly observed that when log-WP was maximized over each fold, the recall reminder was at the optimal level, that is -3.044 for TamperNet, -2.14 for XceptionNet (face), and -3.352 for XceptionNet (Full), respectively.

Korshunov et al. [16] presented DeepFake videos generated from the VID-TIMIT dataset. They used open-source software based on GANs to create DeepFake and assigned special importance to the impact of training and blending parameters on the quality of the resulting low and high visual quality using different tuned parameter sets. They generated two versions of videos for each (320) subject based on low-quality (64×64) GAN model and high-quality (128×128) models. They also demonstrated the SOA on VGG and FaceNet-based face recognition algorithms are vulnerable to the DeepFake videos and fail to distinguish such videos from the original ones with up to 95.00% of false acceptance rate.

Generally, the audio-visual integrated system includes two stages which are to be used for feature extraction and another to classify the tampered videos from the pristine using a two-class classifier. Here, Korshunov et al. [16] followed the same patterns and used Mel-frequency cepstral Coefficient(MFCC) as audio features and distance between mouth landmarks as visual features in the detection pipeline. In the DeepFake videos, the digital presentation attacks consist of Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Image Quality Metrics (IQMs), and Support Vector Machines (SVMs). Dimensionality reduction of the blocks of features for joining the audio-visual

was performed using PCA and then fed to LSTM [17] to separate the tampered and man-tampered videos.

Korshunov et al. [16] also evaluated baseline face-swap detection algorithms and found that the lip-sync-based approach failed to detect mismatches between lip movements and speech. They also verified that image quality measures with a Support Vector Machine (SVM) classifier can detect high-quality DeepFake videos with an 8.97% equal error rate.

Agarwal and Varshney [18] designed a statistical model based on hypothesis testing to detect face-swapped content or fraudulence in images. In this study, the authors considered a mathematical bound value corresponding to the error probability based on the detection of genuine or GAN-generated images.

Lyu [19] highlighted the key challenges in detecting DeepFakes using high-quality or definition-synthesized videos and audio clips. The author tried to raise a deep concern about one of the critical disadvantages of the current DeepFake generation methods that cannot produce a fine mapping of color shades for hair with respect to the human face.

Based on the above discussion, this paper highlighted the laconic view of the proposed DFD pipeline or mechanism and was also concerned about the future amelioration of advanced DFD. Here, the author proposed an adversarial perturbation-enabled model that will give less emphasis on DNN-based face detectors. The proposed detection model consisted of two phases: (a) face detection phase enabling the adversarial perturbation approach and (b) an AI system to detect DeepFake.

Kumar et al. [20] implemented several DL approaches and compared their results with the context of DeepFake classification using metric learning. The authors used a Multitask Cascaded Convolutional Neural Network (MTCNN) to extract faces from images or videos. The MTCNN incorporates three networks: (a) a proposal network, (b) a refine network, and (c) output networks to suppress overlapping boxes using non-max-suppression and generate bounded faces. The Xception architecture was used for transfer learning, and sequence classification was applied using LSTM in addition to 3D convolution and a triple network. A triplet network was used with metric learning for proposing an approach that counts the number of frames in a particular video clip. The realism factor had to be accessed if the number of frames was found to be less than the actual number of frames when compared with pristine video. In this study, three types of triplet-generation methods were investigated. These were easy triplets, semi-hard triplets, and hard triplets, which are based on the distance between the anchor, positive, and negative embedding vectors. The proposed detection architecture, as shown in Figure 7, leverages XceptionNet for the entire process using the MTCNN. In the first phase, the FaceNet model is used to detect, extract, and generate a feature space which is 512 dimension embedding vectors for each face. Subsequently, the generated feature space is fed to semi-hard triplets that discriminate between fake frames and pristine frames through triplet loss. During validation, this approach achieved an AUC score of 99.2% on Celeb-DF and accuracy of 99.71% on a highly compressed neural texture [20].

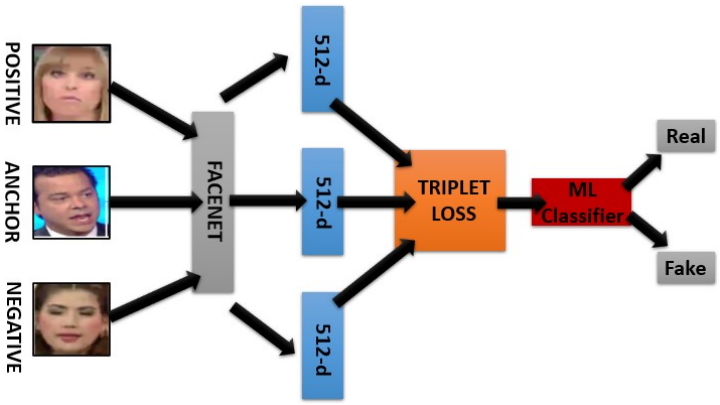

Figure 7. Triplet architecture used for clustering and classification of real videos embedding vectors.

Mittal et al. [21] introduced a method that is a conglomeration of a Convolution Neural Network and a Recurrent Neural Network that helps to extract vital temporal features from faces to detect manipulated or synthesized faces. A Gated Recurrent Unit (GRU), along with a weighting mechanism and Automatic Face Weighting (AFW), was used to automatically choose the most reliable frames for detecting forged faces.

Figure 8 shows the overall execution flow of the proposed detection architecture for knowing the authenticity of genuine or fabricated videos, where this method detects and extracts the facial features from multiple frames using MTCNN. After detecting face regions, a binary classifier is trained using EfficientNet-b5 to extract features that will classify the real and fake faces. Finally, the prediction for classifying realism or fakeness can be estimated by the mixture of AFW and GRU. The authors trained and evaluated the proposed method on the DeepFake Detection Challenge (DFDC) dataset, which yielded a 0.321 log-likelihood error.

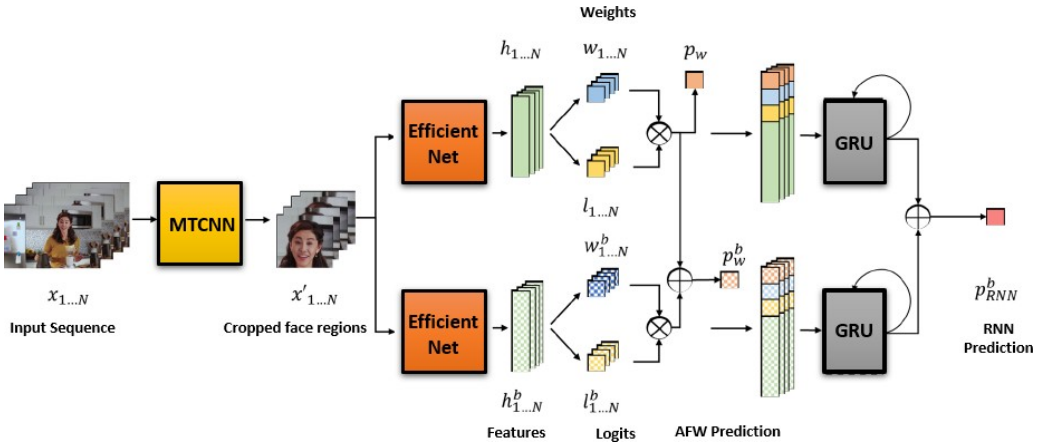

Figure 8. Extraction of the face from the frame using MTCNN algorithm.

Kawa and Syga [22] presented two DeepFake detection models that achieved higher accuracy and low cost in terms of computation. In the first method, they have ameliorated the existing MesNet model by introducing a new activation function, i.e., the Pish activation function. MesNet used a convolution neural network that came in two variants, Meso4 and MesoInception-4. Using MesoNet with the Pish and Mish activation functions showed a higher text accuracy than the other combinations. In the second method, Local Feature Descriptors and the angle of BRISK features were used. In addition, we compared the evaluation performance of the proposed Pish network with that of other benchmark neural networks such as SqueezeNet, DenseNet, and EfficientNet. This method yielded an error rate of 0.28%, which is a comparatively shorter computation time.

Chugh et al. [23] proposed an approach based on the modality dissonance score (MDS), which classifies forgery in DeepFake video between audio and visual modalities through their dissimilarities. The contrastive loss was used to analyze the closeness features between audio and video. In addition, entropy loss is used to analyze the features to detect the individual modality, either audio or video. Kaur et al. [24] used a deep depth-based convolutional long short-term memory model applied to temporal sequential frames to detect DeepFake from the video frames. The frames from the DeepFake video were extracted using "OPENCL", which is generated by a conglomeration of the feature of source frames onto destination frames of the video clip. It is a two-level deep temporal-based C-LSTM, in which the first layer extracts the frame from a forged video and then feeds it to the C-LSTM model for DeepFake detection.

Symeon et al. [25] used the DFDC dataset to train the DL models for detecting fraudulent mages or videos. In this paper, researchers put their efforts into the extraction of face features especially for false positive images, that can generate a large noisy corpus of contents for ameliorating the detection accuracy. Prior to feeding these features to three

proposed deep learning architectures, MesoInception-4, XceptionNet, and EfficientNet, the authors integrated two pre-processing steps: a data augmentation layer and an image filtering layer. In the first step, they pre-processed the dataset, including transformations such as horizontal and vertical flipping, random cropping, rotation, compression, Gaussian and motion blurring, and brightness, saturation, and contrast transformation. This staging layer is used to improve the quality of the image. In the second pre-processing layer, they eliminate images whose sizes are less or equal to N/2 when it is in a connected form, where N is the number of extracted frames per video after the face extraction. At last, the DL models incorporate sigmoid activation in the final layer, Adam optimizer, and minimization in log loss error while training on DFDC. The detailed excerpt of the DeepFake detection pipeline is depicted in Figure 9.

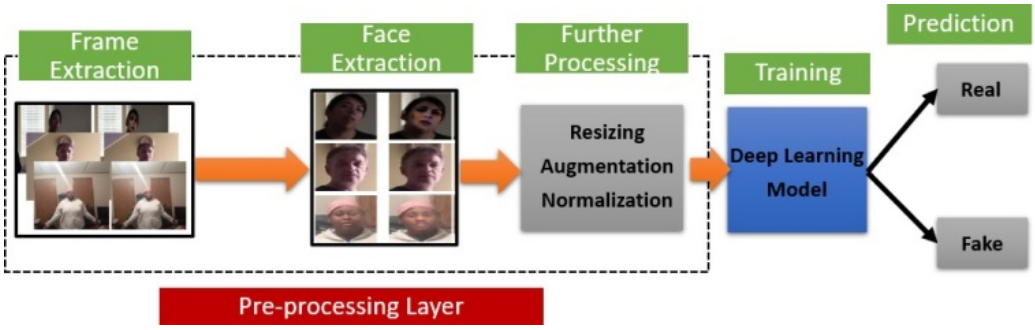

Figure 9. Baseline DeepFake detection pipeline.

Rahul et al. [26] established a technique based on the common attributes of fabricated video clips that analyzed face interpretation. Here, the study consists of a sandwich approach, in which the manipulated videos are converted into frames and fed to the MTCNN to extract the facial features using the MobileNet model. The pre-trained MobileNet is used as an input, and transfer learning is applied to a pre-trained MobileNet neural network to classify the videos as fake or real. This technique was tested on the Face Forensic dataset and obtained an average accuracy of 86% in detection.

Advancements in computer vision and deep-learning methods have led to the discovery of sophisticated and compellingly forged versions of DeepFakes. Owing to the involvement of AI-synthesized DeepFake contents, many attempts have been made to release benchmark datasets and algorithms for DeepFake detection. Prior DFD methods dealt with only a single modality for the authentic originality of videos because researchers have a high error rate in accuracy. In 2020, because of the aforementioned drawback of single-modality analysis, Mittal et al. [21] presented a deep-learning network model inspired by the Siamese network and triplet loss for detecting fake videos. To verify the model, the authors reported the AUC metric on two large-scale DFD datasets: the DF-TIMIT and DFDC datasets. Then, compared with several SOTA DFD methods, such as Two-Stream, MesoNet, HeadPose, FWA, VA, Xception, Multi-task, Capsule, and DSP-FWA, they report a per-video AUC of 84.4% on the DFDC and 96.6% on the DF-TIMIT datasets. This was the first approach to simultaneously exploit audio and video fusion modalities to perceive emotions from a DFD. In this study, the relationship between the visual and audio modalities taken from the same video by extracting the visual face and speech features is shown in Figure 10, which are Freal and Sreal, respectively. Similarly, Open-Face and PyAudio analyses were used to collect meaningful features from visual faces and speech. The extracted features, freal, sreal, ffake, and sfake form the inputs to the networks (F1, F2, S1, and S2, respectively). Later, these networks used a mixture of two triplet loss functions designed using similarity scores denoted as $\rho 1$ and $\rho 2$. Here, $\rho 1$ represents the similarity between the facial and speech modalities, and $\rho 2$ is the similarity between the effects of the modalities of both the original and fabricated videos.

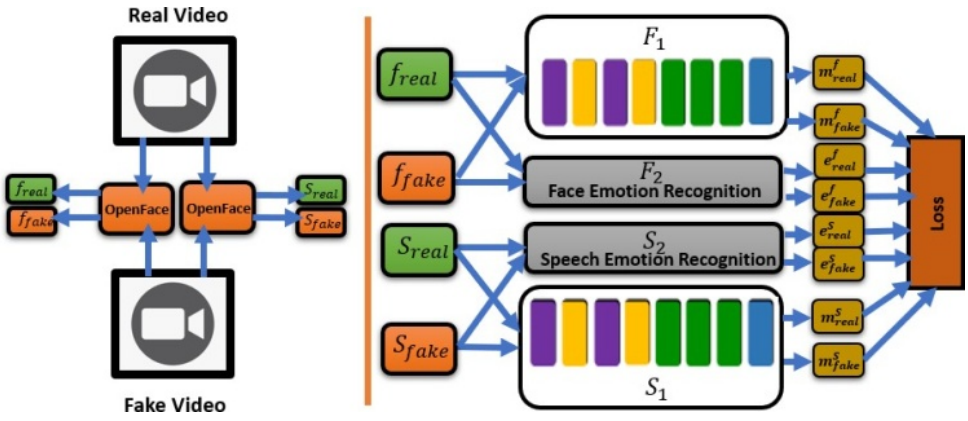

Figure 10. Flow diagram to train a visual-audio detection modelWubet.

Wubet [27] used a ResNet and VGG-16-based CNN to classify eye states and long short-term memory for sequence learning. This study investigated fake or real videos from the UADFV dataset by counting eye blinks within an interval, and the eye aspect ratio was used to determine the height and width of open and closed eyes, as shown in Figure 11.

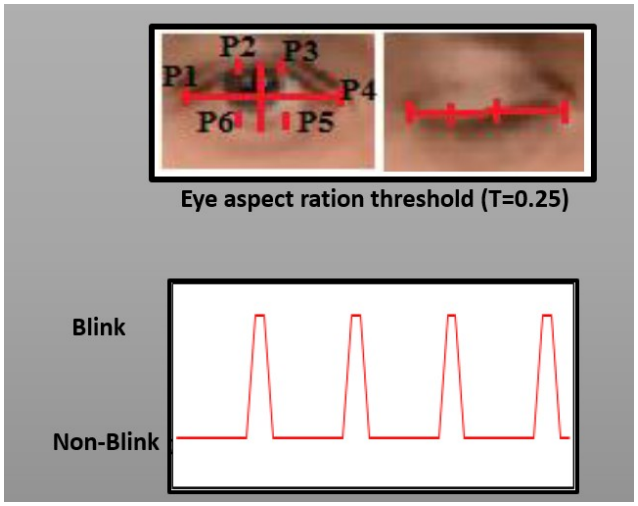

Figure 11. Coordinates to detect eye regions and the blinking of the eye.

Here, in the given figure, p2, p3, p5, and p6 measure the height, whereas p1 and p4 measure the eye width. These points are responsible for determining whether the eyes are closed or open. In this study, the average human eye blinking rate was used as a threshold to detect and count the eye blink and blink intervals because the normal blinking rate of humans is between 2 and 10 s, and each eye blink will take between 0.1 and 0.4 s. Based on this calculation, the authors classified fake and real videos. It detected 184 eye blinks per minute on real videos and 428 eye blinks per minute on fake videos, and the overall accuracy was 93.23% and 98.1% for real and fake videos, respectively.

Similarly, Pishori et al. [28] extended the work on the eye blink mechanism for detecting DeepFakes and proposed a three-stepped detection model which is a combination of convolutional LSTM, eye blink detection, and grayscale histograms to detect DeepFake videos. This study used CNN+RNN integration to detect the number of eye blinks, and the OpenCV library was used to detect facial landmarks in images or video frames.

The entire image pre-processing is performed using grayscale algorithms, and the eye-aspect ratio (EAR) recorded the high accuracy in detection while it is trained. The gray-scale histograms of unaltered and face swap DeepFake are shown in Figure 12.

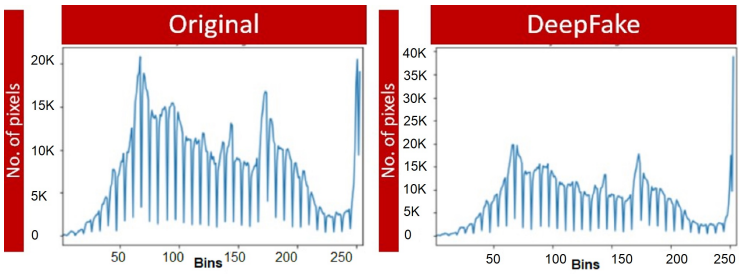

Figure 12. Grayscale histogram of an unaltered video (**left**) and grayscale histogram of a face swap DeepFake of the same video (**right**).

Hussain et al. [29] proposed a novel SOA method for bypassing the DeepFake detector if the adversary has complete or partial knowledge of the detector. In this work, they generated adversarial examples for each frame of a given fake video and combined them to synthesize an adversarially modified video that is classified as real by the victim DeepFake detector, which is XceptionNet and MesoNet, and discussed two pipelines: white-box and black-box attack scenarios to classify the fake videos.

Owing to the increase in the content of DeepFakes, many researchers and reputed universities have already conducted thorough surveys and highlighted the divergent models or systems for detecting the fakeness in images or videos, such as in Vakhshiteh et al. [30]. Nguyen et al. [4], Mirsky and Lee [31], Tolosana et al. [32], Sohrawardi et al. [33], and Verdolina [34] have majorly contributed to highlighting the pluses and minuses of DeepFake and provided a clear excerpt on the mechanisms of DeepFake provided detection.

The rapid increase in research on GANs has led to the generation of finer DeepFake content, which makes it difficult to detect a forgery in videos. Neves et al. [35] implemented five methods on the Celeb-DF v2 dataset using the concept of a spatiotemporal convolutional network to detect DeepFake videos. The authors performed pre-processing on the DeepFake Dataset; at the pre-processing step, the authors cropped the face region from frames, which might distract the overall network learning. The resizing of frames had to be the same size and without any distortion using Retina0Face. Subsequently, it was fed to the respective methods, such as RCN, R2plus1D, I3D, MC3, and R3D, for detecting fake or real videos from DeepFakes. The non-temporal classification method for detecting the originality of frames relies on detecting statistical artifacts in frames generated by the deployment of GANs. The Discrete Fourier Transform (DFT) was applied to the image, and then the 2D amplitude spectrum was compressed into a (300×1) feature vector with an azimuthal averaging mechanism. Afterward, these feature vectors were fed to the classification model i.e., Logistic Regression, which will help to make a decision on the authenticity of the frame or video. The overall architecture of the non-temporal DFD pipeline is shown in Figure 13.

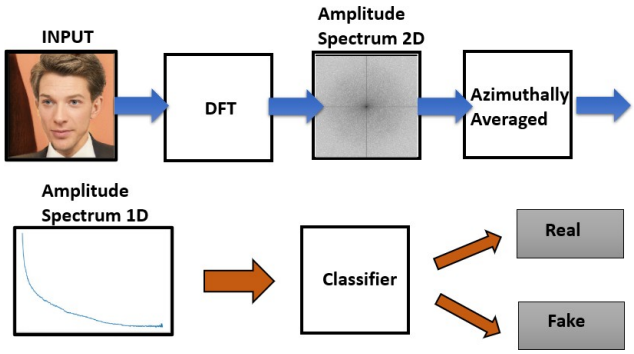

Figure 13. An architecture of non-temporal DFD pipeline.

Guera and Delp [36] proposed a two-stage analysis composed of a Convolutional Neural Network to extract the frame-level features. Later, it had to be fed to a Recurrent

Neural Network that detects the originality of videos or whether they are fake. They achieved precision in accurately capturing temporal inconsistencies between frames due to face swapping. In the CNN stage, Inception-V3 with a fully connected layer at the top of the network was adopted. In the sequential processing stage of LSTM, the SoftMax layer is applied to compute the intra-frame and temporal inconsistencies between frames that are generally created by face swapping or DeepFake manipulation. The LSTM was followed by a 512 fully connected layer with a 0.5% chance of dropout.

Guarnea et al. [37] proposed a technique to analyze the DeepFakes of human faces to detect a forensic trace hidden in images using Expectation Maximization (EM) algorithms. Leveraging EM algorithms are used to extract a set of local features from images, and the validation is performed through tests with a naive classifier on five architectures (GDWCTS, STARGAN, ATTGAN, STYLEGAN, STYLEGAN2) against the CELEBA datasets. Figure 14 shows the overall pipeline of the EM algorithm.

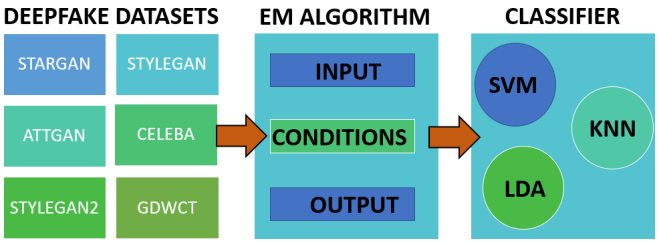

Figure 14. An overview of our attack pipeline to generate Adversarial DeepFakes.

Huang et al. [38] proposed a fake polisher method based on a post-processing shallow reconstruction method without knowing any prior information about the GAN, which can easily fool the existing SOA detection methods. Currently, GAN-based image generation methods are incomplete owing to limitations in leaving some artifact patterns in the synthesized image. Therefore, the authors proposed methods that can easily detect such artifact patterns and reduce the artifacts in the synthesized image. The first trains a dictionary model to capture the patterns of real images and then seeks representation of DeepFake images in a low-dimensional subspace through linear projection or sparse coding. The authors can perform shallow reconstruction of the fake-free version of the DeepFake image, which reduces artifact patterns. Three SOA of DFD methods, namely, GANFingerprint (finger-print based method), DCTA (spectrum-based method), and CNNDetector (image-based method) along with other 16 popular GAN-based fake image generation techniques were used to evaluate whether the images were fabricated or real.

Masi et al. [39] presented a DFD method based on a two-stage network structure that isolated digitally manipulated faces by learning to amplify artifacts while suppressing high-level face content, as depicted in Figure 15.

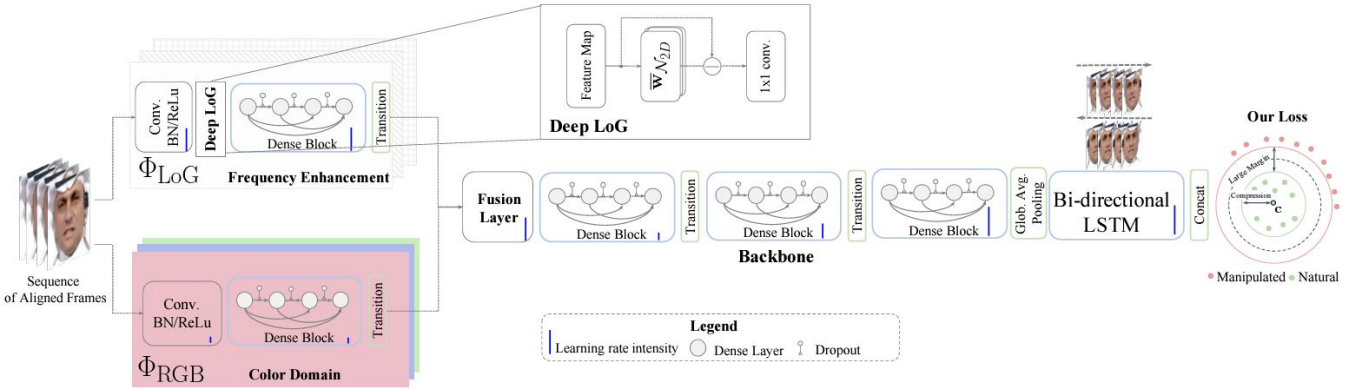

Figure 15. DFD method based on a two-stage network structure.

Currently, this process is using a method that extracts spatial frequencies as a preprocessing step. In this two-branch structure, one branch propagated the original information and another branch suppressed the face content that amplified multi-band frequencies using the Laplacian of Gaussian (LOG) as a bottleneck layer. The LOG operator suppresses the image content presented in low-level feature maps and acts as a band-pass filter to amplify artifacts. The novel loss functions encouraged the compactness of representations of natural faces and provided a way to manipulate faces for better and wider viewpoints. The authors derived a novel cost function for the variability of natural faces and proposed a method for unrealistic facial samples in the feature space. They applied this method to face-forensics, Celeb-DF, and Facebook DFDC-presented benchmarks, as shown in Figure 16.

Figure 16. Unrealistic facial samples in feature space.

The two-branch representation extractor is based on a densely connected layer that learns to combine information from the color and frequency domains using a multiscale Laplacian of Gaussian (LOG) operator.

Trinh et al. [40] proposed a Dynamic Prototype Network (DPNet) that leveraged dynamic representations to explain DeepFake visual dynamics, as shown in Figure 17. The DPNet automatically learns the dynamic prototypes, which can be used to determine the temporal logic specifications that check the robustness of the model and verify whether it is suitable for the desired temporal behaviors. The architecture of the DPNet consists of a feature encoder, prototype layer, and temporal logic for verifying dynamic prototypes. The evaluation of these methods was made on the DFD and FF++ datasets and tested with other baseline models, as shown in Figure 17. In this study, the authors used a quantitative interpretation metric to measure the interpretation against the ground truth to determine how well the prototypical patch of the prototype overlapped with the ground truth mask.

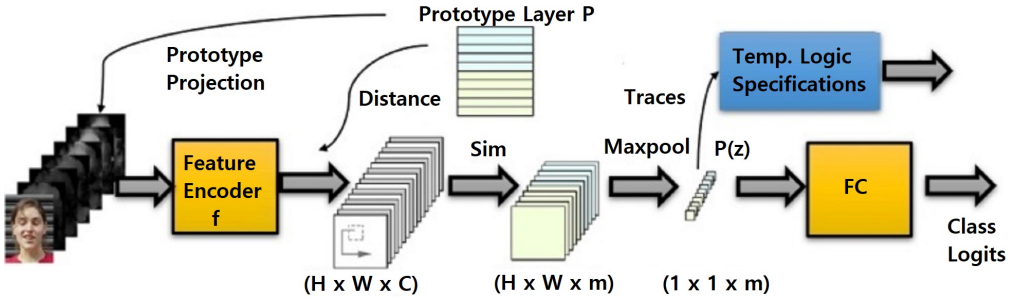

Figure 17. Proposed DPNet architecture for detecting fake images.

Li et al. [41] proposed a novel method called Face X-ray to detect a forgery in face images and provide blending boundaries of a forged face using a binary mask. The detailed architecture of the detection method is depicted in Figure 18, in which the authors blended the altered face into an existing background image, and intrusive image discrepancies existed across the blending boundaries.

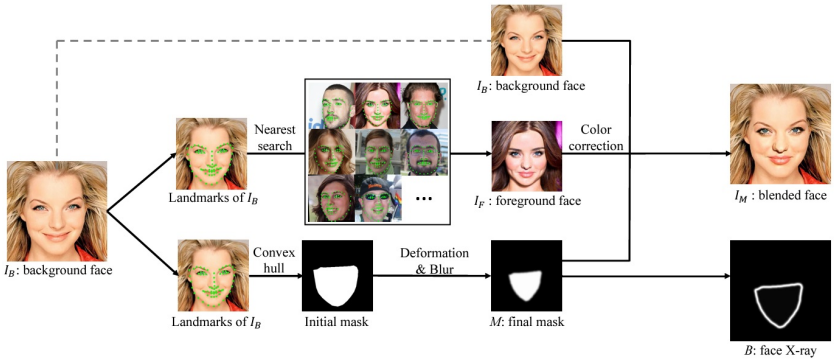

Figure 18. Detailed description of Face X-ray detection architecture.

The grayscale image acts as an input face image for the Face X-ray method, which reveals whether it can be decomposed into a blend of two images from a disparate source. During testing, the binary face boundaries will be generated with a forgery in the image owing to blending, and it generates a blank image when the image is real. Owing to the acquisition process, each image has its own distinctive marks, that is, noise and error level analyses of distinctive marks, as shown in Figure 19.

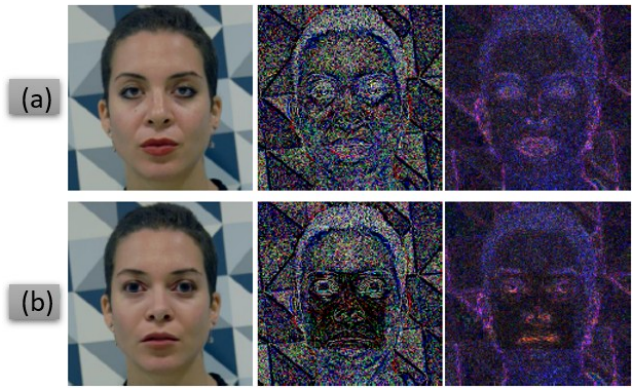

Figure 19. Noise analysis (middle) and error level analysis (right) of (**a**) a real image and (**b**) a fake image.

Li et al. [42] extended their previous work on facial X-rays. In this study, they examined a novel face-swapping algorithm called a face shifter for high-fidelity and occlusion-aware face swapping, as shown in Figure 20.

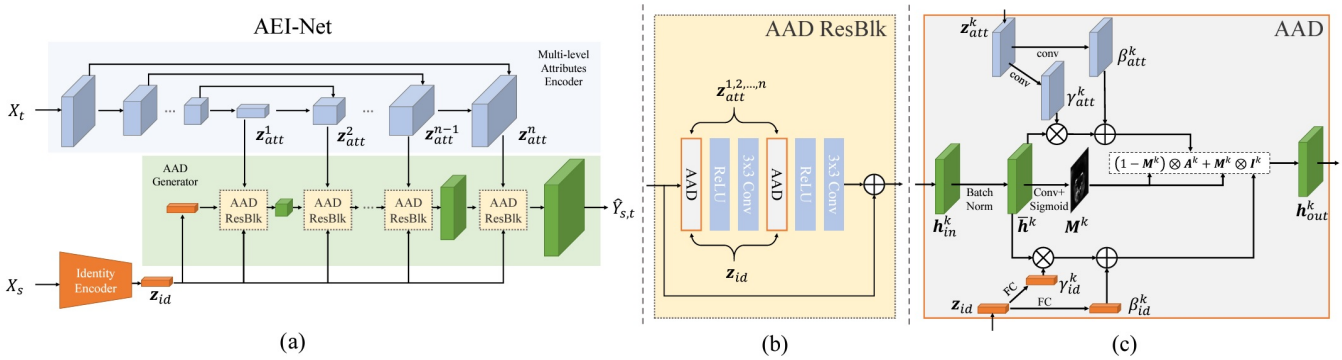

Figure 20. AEI-Net is composed of an Identity Encoder, a Multi-level Attributes Encoder (**a**), and an AAD-Generator. The AAD-Generator uses cascaded AAD ResBlks (**b**), which are based on AAD layers (**c**), to combine identity and attribute information at different feature levels [42].

The face shifter generated a swapped face with high fidelity by exploiting and integrating target attributes. It can handle faced occlusions with a second synthesis consisting of a heuristic error-acknowledging reinforcement network (HEAR-Net). They also developed a Face X-ray method to detect forged images created by a face shifter. The authors designed a GAN-based network called adaptive embedding integration networks (AEI-Net) for the integration of target attributes and multilevel attribute encoders instead of compressing it into a single vector, such as RSGAN and IPGAN, in addition to AAD layers. Such adaptive integration improved the signal-level integration used by the RSGAN, FSNet, and IPGAN.

Neves et al. [35] proposed a GAN-fingerprint removal approach (GANprintR) to generate a more realistic synthesis or DeepFake Dataset that can be used by a research group for the detection of DeepFakes, as shown in Figure 21. In this study, the authors used three detection models or classifiers based on XceptionNet, Steganalysis, and local artifacts to check for fake detection on the synthesized dataset. When tested with XceptionNet, it obtained an average absolute worsening of 9.65% EER when using GANprintR. However, the degradation was higher for Steganalysis (14.68% EER) and more promising with local artifacts (4.91% EER).

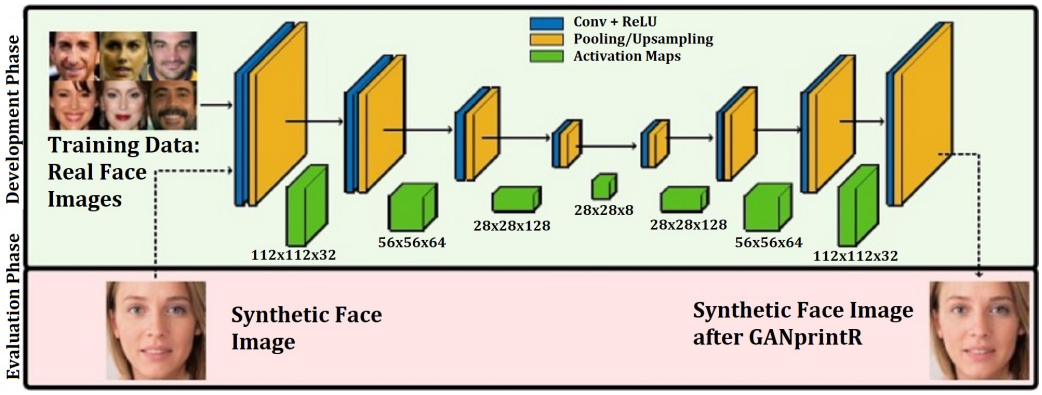

Figure 21. Proposed video-based face manipulation detection architecture.

In more recent years, Yang et al. [43] further provided a new perspective by formulating DeepFake detection as a graph classification problem, in which each facial region corresponds to a vertex. A spatiotemporal attention module was exploited to learn the attention features of multiple facial regions with improved DeepFake detection.

Zhao et al. [44] proposed an Interpretable Spatial-Temporal Video Transformer (ISTVT), which consisted of a novel decomposed spatial-temporal self-attention and a self-subtract mechanism to capture spatial artifacts and temporal inconsistency for robust Deepfake detection. The authors proved improved DeepFake detection performance in extensive experiments using large-scale datasets, including FaceForensics++, FaceShifter, DeeperForensics, Celeb-DF, and DFDC datasets.

Wang et al. [45] proposed a deep convolutional transformer to incorporate the decisive image features both locally and globally. The authors applied convolutional pooling and re-attention to enrich the extracted features and enhance efficacy. Their work employed image keyframes in model training for performance improvement and visualized the feature quantity gap between the key and normal image frames caused by video compression with improved transferability in DeepFake detection.

Yu et al. [46] proposed a novel Augmented Multi-scale Spatiotemporal Inconsistency Magnifier (AMSIM) with a Global Inconsistency View (GIV) and a more meticulous Multi-timescale Local Inconsistency View (MLIV), focusing on mining comprehensive and more subtle spatiotemporal cues. Yu et al. [47] further used Predictive Visual-audio Alignment Self-supervision for Multimodal DeepFake Detection (PVASS-MDD), which consisted of PVASS auxiliary and MDD stages for DeepFake detection. The continued advancement in spatiotemporal and multi-modal fusion in DeepFake detection is noted using advanced machine learning techniques.

4. Audio Modality Fusion in DeepFake Detection

Similar to DeepFakes in video and image, audio content can be a victim of DeepFakes (and also contribute to hyper-realistic DeepFakes on video data). The increase in fake and synthetic audio has become one of the major challenges for researchers in distinguishing between spoofed audio and genuine audio. The first well-publicized instance of an audio DeepFake scam was reported in mid-2019. At that time, the fraudsters used artificial-intelligence-based software to impersonate a chief executive's voice and demand a fraudulent transfer of USD 243,000 [48]. Automatic speaker verification (ASV) systems are primarily threatened by replay and audio spoofing attacks, voice conversion (VC), and speech synthesis (SS) to commit illegal acts.

SS and VC have also progressed significantly over the past decade, reaching a point where it has become very challenging to differentiate between spoofed speech and genuine user speech. Technical enhancements in synthetic audio and DeepFakes threaten to magnify the scale, persistence, and consequences of misinformation. Incorrect information can affect emotions and opinions. At worst, it could lead to organized and stabilized unwanted public actions united behind false intentions or impressions. To avoid SS and VC attacks, several researchers have adapted the ASV technique with audio spoofing detection systems that feature countermeasure scores to classify spoofed and genuine speech. The ASVSpoof challenge edition was initiated, as shared in Table 2.

Table 2. DeepFake classification based on Audio.

Dataset	Total Audio	Release Date
AVSSpoof2015	106	17 September 2015
AVSSpoof2017 V2	179	2 April 2018
AVSSpoof2019	107	4 June 2019

Chettri et al. [49] studied the state of the art and observed the model speaker performance in an end-to-end manner for the ASVSpoof2017 challenge. They found that architectures such as the second version of the automatic verification spoofing and countermeasures challenge (ASVSpoof2017) showed poor generalization in the evaluation dataset but found a compact architecture that showed good generalization on the development data, which demonstrated that it was not easy to obtain a similar level of generalization on both development and evaluation data, leading to a variety of open questions. Here, the authors reported their experiments and challenges in designing a deep anti-spoofing system that was trained and evaluated on the ASV spoof database. They explored four end-to-end CNN-based models generalized in the development dataset, but consistently performed well in the evaluation dataset. Later, they explained their experiments to determine a suitable architecture that generalizes well to unseen data. They also proposed a novel CNN architecture for the spoofing detection task that has approximately 5K free parameters, as shown in Figure 22, in which the shape of the feature map after the second convolutional an max pooling layer is ($8 \times 25 \times 33$), that is, no. of channels $\times$ time $\times$ frequency.

Tom et al. [50] proposed a detection method consisting of a group delay gram (GD-gram) obtained by concatenating a group delay function over consecutive frames as a novel time-frequency representation of an utterance. This work was divided into two stages. The first stage incorporates transfer learning of a pre-trained CNN for fast adaption to the GD-grams extracted from utterances and attentional weighting of the raw GD-grams is performed; in the second stage, another stage of transfer learning of a pre-trained CNN on GD-grams weighted by soft attention for classification is performed. The model adapted the ResNet-18 architecture and used its Global Average Pooling (GAP) layer to provide attention maps for the second stage of discriminative training to improve performance. Figure 23 shows the overall detection framework for identifying spoofs in audio.

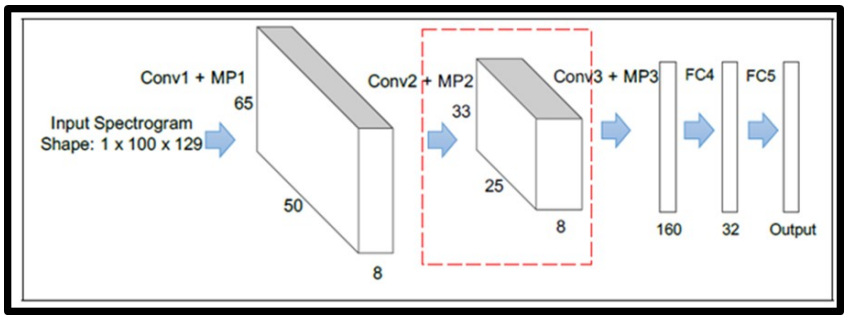

Figure 22. Architecture of the proposed model, where Conv: convolutional layer, FC: fully connected layer, MP: max pooling layer [49].

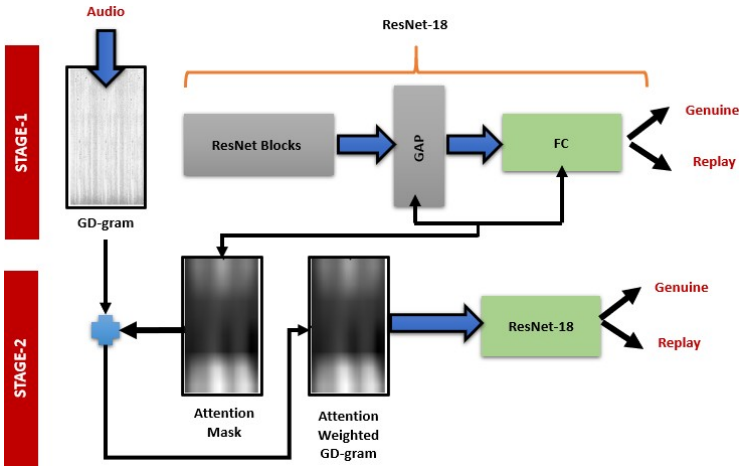

Figure 23. Proposed audio-based spoof detection architecture [50].

Alzantot et al. [51] developed an audio spoofing detection system based on the countermeasure score (CMS) to distinguish between spoofing attacks and pristine speech. In this paper, authors incorporated three variants of feature extraction: the Mel-Frequency Cepstral Coefficient (MFCCS), the constant Q Cepstral Coefficient (CQCCS), and the Logrithmic Magnitude of Short-Time Fourier Transform (Log-magnitude STFT). The other three deployment models, MFCC-ResNet, CQCC-ResNet, and Spec-ResNet, along with their respective feature extractors, were applied to 78 human voice clips for evaluation. Finally, they evaluated the performance of RCN with varying choices of input features against the two attack scenarios of ASVSpoof2019 (logical access and physical access) using both the development (known attacks) and evaluation datasets (both known and unknown attacks). The outcome of this model showed an overall improvement in t-DCF and EER scores of 71 and 75, respectively.

Todisco et al. [52] addressed two key novel advancements in detecting spoofs in audio. These advancements include (a) addressing two different spoofing scenarios, Logical Access (LA) and Physical Access (PA), along with their three important forms of spoofing attacks, that is, synthetic, converted, and replayed speech, and (b) use of the tandem detection cost function (t-DCF), which reflects the impact of both spoofing and countermeasures on ASV reliability. Logical Access aims to determine whether the advances in TTS and VC technology pose a greater threat to the reliability of ASV scenarios. The Physical Access scenarios aimed to assess the spoofing threat and countermeasure performance via simulation, with which factors influencing replay spoofing attacks could be carefully controlled and studied. The PA database was constructed from a far more controlled simulation of replay spoofing attacks, which are also relevant to the study of fake audio detection in the care of smart home devices. ASVSpoof 2019 migrates to a new primary metric in the form of an ASV-centric tandem decision cost function (t-DCF). The adoption

of t-DCF ensures that the scoring and ranking reflect the comparative impact of spoofing and countermeasures on an ASV system.

Balamurali et al. [53] showcased a detailed implementation step of SOA that consisted of both the pre-processing and post-processing of audio signals or speech to detect the originality of audio content. In this study, the authors focused on an ensemble method by considering the correlation between several audio spoofing detection techniques. In this spoof detection system, they examined all robust audio features, including traditional and learned features, using an autoencoder. The base layer for implementing this system uses a traditional Gaussian mixture model called the universal background model (GMM-UBM). When evaluated on the ASVSpoof2017 database, this feature ensemble model showed an equal error rate (EER) of 12, which was further improved to 10.8 by introducing a hybrid model with the conglomeration of both known and machine-generated features that are trained on an augmented dataset.

Kamble et al. [54] made a great effort to investigate the Teager energy-based features for spoof speech detection (SSD) tasks. The Teager energy profiles computed for natural, VC, SS, and replay signals showed changes around the glottal closure instants (GCIs). For the SS signal, the bumps were very smooth compared with the natural signal. These variations around the GCI of the Teager energy profiles helped discriminate the spoof signal from their natural counterparts. The Teager energy-based feature set, that is, Teager Energy Cepstral Coefficients (TECC), performed outstanding for S1–S9 spoofing algorithms with an average EER of 0.161%, whereas state-of-the-art features, namely, Cochlear Filter Cepstral Coefficients-Instantaneous Frequency (CFCC-IF), and Constant-Q Cepstral Coefficients (CQCC) gave an EER of 0.39% and 0.163%, respectively. It is interesting to note that the significant negative result of the proposed feature set to S10 vs. natural speech confirms the capability of TECC to represent the characteristics of airflow patterns during natural speech production. Furthermore, the experiments performed on the BTAS 2016 challenge dataset, gave 2.25% on the development set. In the evaluation set, the TECC feature set gave a Half Total Error Rate (HTER) of 3.8%, which is the metric provided by the challenge organizers, thus overcoming the baseline by a noticeable difference of 3.16%. However, the TECC feature failed to detect the USS-based spoofing algorithm and unknown attack of the replay speech recorded with the laptop HQ device. The error rate of TECC+MFCC is 0.38% on the development set and 6.41% on the evaluation set.

Chen et al. [55] proposed an audio-based model to detect audio DeepFakes or limit spoofing of voices, and extracted 60-dimensional linear filter banks (LFBs) from raw audio and passed them into a residual network. FreqAugment and augment layer and large margin cosine loss function (LMCL) were being used during the training. The main objective of LMCL is to maximize the variance between genuine and spoofed class and FreqAugment, a layer that randomly masks adjacent frequency channels while in DNN training for increasing the generalization ability of the DNN model.

Kumar and Bharathi [56] proposed a novel feature called the filter-based cepstral coefficient (FBCC) 55, which is used in the front-end processing of countermeasures. FBCC is a new feature extraction approach that was proposed and used for the first time in the field of speech processing, particularly for spoof detection in ASV systems. The FBCC-based countermeasure was substantially effective in counterattacking spoofed utterances under both the LA and PA conditions. FBCC is based on the energy variation pattern (EVP), which captures the energy variation information with regard to the energy level in the neighborhood within a frame and adjacent frames. EVP is computed using statistical filters; hence, the proposed approach is called the filter-based cepstral coefficient (FBCC). Three versions of the FBCC were analyzed in this study. The different versions of the FBCC are based on the types of filters used, namely Gaussian, bilateral, and median filters. The computational similarity between the FBCC with linear frequency cepstral coefficients (LFCC) and Mel frequency cepstral coefficients (MFCC) has led to the consideration of MFCC and LFCC as baseline countermeasure systems. The main strengths of the FBCC were 1. generalization of spoof detection for different types, 2. better performance under

LA and PA conditions, and 3. the fine-tuning of the parameters in the context of the filters was comparatively lower.

The computation of the FBCC uses the power spectral density (PSD) as the basis. The PSD is a function of the energy strength in the frequency domain. The FBCC tends to capture the changing pattern in energy relative to the changes in energy strength in the frequency domain. The FBCC captures energy variations intended to discriminate synthetic speech from natural speech.

Chintha et al. [57] proposed a new audiovisual authentication detection method using a combination of convolutional latent representations with bidirectional recurrent structures, such as CRNNSpoof and WIRENetSpoof, and entropy-based cost functions. Latent representations for both audio and video were carefully chosen to extract semantically rich information from the recordings. By feeding these into a recurrent framework, we can detect both the spatial and temporal signatures of DeepFake renditions. Entropy-based cost functions work well in isolation and in the context of traditional cost functions. They demonstrated the methods on the FaceForensics++ and Celeb-DF video datasets and ASVSpoof 2019 Logical Access audio datasets, thereby achieving new benchmarks for all categories. They performed extensive studies to demonstrate generalization to new domains and gain further insight into the effectiveness of the new architectures. These audio embeddings were passed into a bidirectional recurrent layer.

Das et al. [58] conducted a comprehensive analysis of the nature of different types of spoofing attacks and system development, particularly long-range acoustics and deep features for spoofing detection. In the training phase of the deep feature extractor (DFE), it incorporated a discrete Fourier transform (DFT) and fed to the log power spectrum process as input from attributes related to utterances that considered both bonafide and spoofed speech from the train set, which can be used as DFE by eliminating the output layer and then, generating the embedding as a deep feature representation.

5. Advanced DeepFake Detection Methods

Li et al. [59] focused on the advancement of facial landmark detection algorithms and improved picture and video manipulation techniques and the integration of generative models such as GANs and VAEs, which has contributed to more convincing and realistic DeepFakes. Thereafter, Cozzolino et al. [60], in 2021, demonstrated the application of DeepFake in forensic investigation, machine-learning-based classification models, and the examination of visual artifacts and inconsistencies. The temporal aggregation of convolutional representations and deep learning techniques were also mentioned as having demonstrated promising results in the detection of DeepFakes.

Zaho et al. [61] drew attention to the growing danger posed by DeepFake videos, which are realistic but artificially produced videos that might trick viewers by depicting things or people that do not actually happen or exist. Owing to their increasingly complex generating processes, modified films are sometimes difficult for traditional DeepFake detection systems to recognize correctly. The authors suggested a multi-attentional strategy that combines self-attention, spatial attention, and temporal attention mechanisms to overcome this difficulty. These attention processes enabled the model to focus on essential regions and patterns while filtering out extraneous data, allowing it to effectively capture both global and local contextual information within videos.

The proposed DeepFake detection model can identify artifacts, inconsistencies, or anomalous patterns that point to DeepFake manipulation by incorporating these attention mechanisms. For decision-making, the model examines visual and temporal clues, such as facial expressions, eye movements, and motion patterns. The authors stressed the value of using sizable datasets that include a variety of DeepFake variations while training the multi-attentional DeepFake detection model. The generalization and resilience of the model against unknown manipulation approaches were enhanced using this method. To excel on various DeepFake video formats, the model can also benefit from transfer learning and domain-adaptation techniques. However, the authors noted that there is still competition

between DeepFake production techniques and detection approaches. To remain ahead of harmful actors and ensure the development of efficient DeepFake detection techniques, they emphasized on the necessity for ongoing research, innovation, and collaboration among the scientific community, industry, and governments.

Zhou et al. [62] highlighted that, while visual cues have been extensively utilized in DeepFake detection, audio information can provide valuable complementary signals. Manipulated videos often exhibit discrepancies between audio and visual components because of the challenges of synchronizing fake audio with manipulated visual content. To address this, the authors proposed a joint audio-visual DeepFake detection approach that simultaneously analyzes both audio and visual aspects of videos. The model leverages deep learning techniques to extract relevant features from both modalities and integrates them to make a joint decision regarding the authenticity of the video. The visual component of the model utilizes Convolutional Neural Networks (CNNs) to extract visual features from frames or facial regions of the input video. These features capture visual cues such as facial expressions, inconsistencies in facial movements, or artifacts introduced during DeepFake manipulation.

Simultaneously, the audio component of the model employs audio processing techniques, such as spectrogram analysis, to extract relevant audio features. These features capture acoustic cues such as speech patterns, speaker characteristics, and anomalies in audio quality. The extracted audio and visual features are then fused using fusion mechanisms, such as concatenation or attention mechanisms, to create a joint representation that captures combined information from both modalities. This joint representation is fed into a classification model that determines whether the video is genuine or manipulated.

The authors emphasized the importance of training the joint audio-visual DeepFake detection model on diverse datasets that include a wide range of DeepFake variations. This enables the model to learn discriminative patterns and generalize well for unseen manipulation techniques. The experimental results presented in their paper demonstrate that the proposed joint audio-visual DeepFake detection approach outperforms the individual audio-only or visual-only approaches. The fusion of audio and visual modalities leads to improved detection accuracy and robustness against various DeepFake manipulation techniques. Zhao et al. [63] became aware of the growing danger posed by DeepFake films and the demand for effective detection techniques. Traditional methods frequently rely on a single characteristic or modality, which may limit their ability to identify complex DeepFakes. They suggested MTFF-Net, a multi-feature fusion network, as a solution to this problem.

MTFF-Net used a variety of visual elements retrieved from DeepFake videos to improve detection. Color histogram, optical flow, Convolutional Neural Networks (CNNs), and long short-term memory (LSTM) features were the four main visual features of the network. Each element served as a representation of a different aspect of the video content and offered helpful hints for differentiating between real and fake videos.

The color histogram feature recorded statistical data on color distributions in frames, allowing for the detection of anomalies or inconsistencies caused by DeepFake manipulation. Using the optical flow function, it was possible to identify anomalies that may be present in DeepFake videos by capturing the motion patterns between frames. The CNN features were extracted using pre-trained CNN models, which selected high-level representations from the input frames. These features could distinguish between real content and staff that have been altered, and record intricate visual patterns.

The LSTM features were obtained from LSTM networks, which considered the sequential data between frames and captured the temporal dynamics of the video. These characteristics were particularly helpful for spotting temporal irregularities that DeepFake videos frequently contained. The authors' multi-branch architecture extracted discriminative representations from each modality by processing each piece of visual information separately. The features were then combined at several levels, enabling the network to take advantage of the complementary data offered by each feature. The authors employed a

sizable dataset that contained a wide variety of DeepFake movies to train MTFF-Net. To optimize the network parameters and facilitate precise detection, they used proper loss functions and optimization approaches. The experimental findings in this study showed that when compared to single-feature-based approaches and other cutting-edge DeepFake detection models, MTFF-Net performed better. The network can capture a thorough grasp of the video content owing to the multi-feature fusion strategy, improving the detection accuracy and robustness against various DeepFake manipulation approaches.

In 2022, Varma and Rattani [64] focused on the development of a gender-balanced DeepFake dataset specifically designed for FIR (face-in-video) DeepFake detection. The key contribution of this study is the introduction of the GBDF dataset, which aims to address the gender bias commonly observed in existing DeepFake datasets. Gender bias refers to an imbalance in the representation of males and females in the dataset, which can result in a biased performance of DeepFake detection models. The GBDF dataset is designed to have an equal number of male and female subjects and includes a diverse range of facial expressions, lighting conditions, and camera angles. It may contain both genuine and DeepFake videos, with DeepFakes generated using various manipulation techniques, such as face swapping or facial reenactment.

This paper discussed the process of collecting and curating the GBDF dataset, ensuring that it represents a comprehensive and balanced dataset for DeepFake detection research. It may include details of the annotation process, data preprocessing steps, and any specific challenges or considerations in building a gender-balanced DeepFake dataset. Furthermore, this paper presents experimental evaluations using the GBDF dataset to demonstrate its effectiveness in training and evaluating DeepFake detection models. This could involve comparing the performance of models trained on GBDF with those trained on other existing datasets, highlighting the benefits of gender balance in improving the detection accuracy and robustness.

In the same year, Jia et al. [65] developed a face forgery detection method that utilized a fusion of global and local features. Face forgery detection aims to identify instances in which a person's face has been manipulated or replaced with another person's face. Global features typically refer to holistic characteristics that capture the overall appearance and structure of a person's face. These features may include facial landmarks, color distributions, texture patterns, and statistical information. Global features provide a high-level understanding of the face and can help identify inconsistencies or anomalies introduced by facial forgery.

Local features focus on specific regions or patches within the face. These features capture fine-grained details such as textures, edges, and local patterns. By analyzing local features, this method can detect subtle discrepancies or artifacts that may be indicative of face manipulation or forgery. The GLFF approach combines global and local features to leverage complementary information. The fusion of global and local features aims to enhance detection accuracy by capturing both the overall structure of the face and local details. Finally, this study may include experimental evaluations to assess the performance of the GLFF method. This could involve testing the approach on benchmark datasets containing genuine and manipulated facial images or videos. The evaluation may measure metrics such as accuracy, precision, recall, and F1 score to demonstrate the effectiveness of the GLFF method compared with existing approaches.

In 2023, Yan et al. [66] proposed a method called Uncovering Common Features (UCF), which focuses on identifying common visual patterns and features across different DeepFake manipulation techniques, leading to more robust and generalizable detection models. The UCF method aims to uncover the common visual patterns and features shared by different types of DeepFake videos. By focusing on these shared features, the approach aims to develop a more generalizable detection model that can effectively detect a wide range of DeepFakes. The UCF method employs deep-learning architectures to extract discriminative features from DeepFake videos. The features were designed to capture both global and local characteristics and provide a comprehensive representation of the videos.

Mcuba et al. [67] explored the impact of deep learning methods on the detection of DeepFake audio, specifically in the context of digital investigations. The authors investigated different deep-learning techniques and their effectiveness in identifying manipulated and synthetic audio content. Their research focused on various deep-learning methods and architectures employed for DeepFake audio detection. These may include Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), generative adversarial networks (GANs), and other deep learning architectures that have been used for audio analysis and classification tasks. The authors likely curated a dataset specifically designed for DeepFake audio detection. This dataset may consist of both genuine and manipulated audio samples, representing a range of DeepFake audio techniques such as speech synthesis, voice conversion, and audio manipulation.

Evaluating the impact of various data augmentation techniques on the training of DeepFake detection systems helps improve precision and robustness against manipulated content. There are several techniques [68]:

- Flipping
- Color space
- Cropping
- Rotation
- Translation
- Noise injection
- Color space transformations
- Kernel filters
- Mixing images
- Adversarial training
- GAN-based data augmentation
- Neural style transfer
- Meta-learning data augmentation

Generative AI, including DeepFake technology, is highly susceptible to adversarial attacks due to its "neural-network-based" nature. Adversaries can exploit vulnerabilities to generate deceptive content that may bypass detection. Enhanced adversarial training, incorporating diverse adversarial examples during model training, and deploying defensive mechanisms such as adversarial loss can improve resilience against intentional manipulations [69].

6. Conclusions and Future Scope

The image and video feature sections on DFD elicited readers to become familiar with all the novel efforts that have been made by researchers from late 2017 to date. Although the work conducted on DeepFakes by the researchers or research groups has indeed progressed a lot towards the refinement and betterment in the existing models, there is still a large scope of further research for improving the detection pipeline in terms of precision, time efficiency, cost efficiency, and ease of interaction with real-world applications, which can curtail and act as fuel for this DeepFake detection challenge. DeepFake detection models often struggle to generalize across diverse datasets, leading to reduced effectiveness in real-world scenarios with variations in lighting conditions, facial expressions, and video quality. Another main challenge is raised due to the "unseen class of some facial datasets" in the testing dataset with respect to the training dataset. Augmenting training datasets with diverse samples, employing transfer learning from pre-trained models, and integrating attention mechanisms can enhance generalization capabilities [45]. Future research topics can include:

- Investigating the role of kernel dimensions when extracting features through EM algorithms.
- Evaluation of different DFD techniques using real and manipulated datasets, including full-body DeepFakes.

- Efficient, reliable, cross-platform robust mobile applications to detect DeepFake images and videos. Employing platform-agnostic frameworks and optimizing model architectures for mobile devices ensures accessibility and usability across various platforms. Consideration for resource constraints on mobile devices is crucial.
- Leveraging model pruning, quantization, and efficient training techniques can optimize precision, time efficiency, and cost efficiency, making DeepFake detection pipelines more practical for real-world applications [70].
- Integrating temporal logic specifications into detection models enhances interpretability and helps capture temporal patterns indicative of DeepFake content, providing more context-aware detection
- Adopting big data architectures or in-memory distributed frameworks can improve computation efficiency and facilitate real-time DeepFake detection. However, challenges include data management and system complexity. Frameworks like Apache Spark, and Apache Kafka exemplify one of the real-time frameworks for DeepFake detection. But still, it requires continuous maintenance of the queuing system and persistence layer. On the flip side, SaaS-based services have flexibility to easily deploy and monitor the DeepFake detection pipeline, but it does require high configuration cloud instances, and ultimately increasing the cost and portability from one system to another system is non-trivial. Hence, another research domain might be influenced towards edge computing (decentralized computing) for impeccable and precise real-time frameworks [71].
- Implementing the fusion of different modalities by creating a correlation mechanism among several results of DFD methods for better performance.
- The incorporation of different data augmentations prior to training of the DFD system can improve the precision in detecting whether the image or video is pristine or fake.
- Focus on expanding the LFD-based technique to achieve a lower EER, in addition to less time and computation of DFD.
- Incorporating temporal logic specifications can increase the scope of interpretability.
- Use of distributed computing or distributed lightweight virtual machines to support real-time detection systems.
- To extend the use of the unsupervised domain, the feature space from the source dataset is adapted to the target dataset to make the model robust and label-independent.
- Exploring lightweight model architectures, model compression techniques, and edge computing solutions can mitigate resource constraints. Optimization of algorithms and prioritizing essential features can reduce resource requirements without compromising detection accuracy. Alternatively, knowledge distillation can be beneficial to alleviate the hassle of high computation for deploying a DeepFake detection algorithm by leveraging the teacher–student architecture.
- Continual unsupervised learning can be used to manage the resource-intensive nature of advanced ML or DL techniques, but it is prone to catastrophic effects. Further research scope is to implement the "prompt" in the continual learning for DeepFake detection.

In audio DeepFakes, there are further futuristic scopes to improve the improve the accuracy of authenticating spoofed or genuine audio. Some main challenges analyzed during the survey are listed below:

- Researchers have suggested improving the generalization of the model against unknown spoofing attacks by applying advanced fusion to build a "wide and deep" network that concatenates the features of the last fully connected layers of each model with a shared soft-max layer as the output layer to ameliorate the fusion result.
- Researchers are focusing on Unit Selection Synthesis (USS)-based spoof detection, which is a festival framework that compromises different modules, such as Text Processing, Phonetic Analysis, Prosodic Analysis, and Speech Generation.

- Investigating multi-modal fusion techniques, such as attention-based fusion and graph neural networks, can improve the integration of information from different modalities for more robust DeepFake detection.
- Extending unsupervised domain adaptation techniques enhances the robustness of DeepFake detection models when applied to new and diverse datasets without labeled samples. For adapting the unsupervised domain in the DeepFake, various researchers have been working towards the Zero-Shot Learning (ZSL), Few-Shot Learning (FSL), and attention-based Online Transfer Learning (OTL).
- Lastly, the researchers are willing to further investigate bank-of-classifier solutions to detect spoofness attacks that also require different solutions or fusions.

As DeepFakes can have significant impacts on society overall, by establishing platforms for collaboration between government agencies, industry stakeholders, and research institutions fosters a collective effort to address the societal impact of DeepFakes effectively, sharing insights and resources. Currently, the governments of several countries have started concerning the negative impact of the DeepFake that creates a room with scope of further improvement in terms of research.

In DeepFakes, there are a myriad of transformation approaches to explicitly endow the perturbations for generating the falsified images or videos that seem realistic to humans. There are mainly three methods to generate and detect the DeepFake by leveraging machine learning, deep learning, and rule-based learning to deep scan the extracted spatial and temporal features as embeddings and analyze if any attacks exist in the image or video. Moreover, the most common transformations for generating the DeepFakes are face swapping, lip-syncing, expression alteration, hair alteration, false gestures, fake speech synthesis, background replacement, and eye-gaze manipulation. Furthermore, falsification might occur due to soft/unseen attacks, which are difficult to see and analyze with lightweight detection models or normal investigation, and hard/seen attacks can be visible by humans or lightweight detection models. If we accumulate all the transformations, it would be a non-trivial problem to design robust and generic algorithms to detect all kinds of perturbations in the image or video because some algorithms are specifically designed to work only for the image dataset and others for both audio and visual. In other words, some attack detection methods are designed and trained by considering some specific transformations, while the others are targeting to detect different transformations. Furthermore, the performance of detection models might be degraded when the testing set is different from the training set with significant diversity. Thus, it requires further research on improved model generalization for designing robust and cross-dataset-adaptable models.

Zhang et al. [72] observed that several works have been performed on DeepFake detection of attacks or any other adversarial noises towards images or videos that might not be adaptable to the detection of audio transformations. In a nutshell, no detection method is completely immune to adversarial attacks. Many researchers and practitioners continuously work on improving detection methods and developing robust defenses against adversarial attacks by using the concept of unsupervised learning as well.

Continuous monitoring of DeepFake generation techniques is crucial. Leveraging unsupervised learning approaches, anomaly detection, and real-time model updates can aid in adapting to emerging forms of synthetic media.

Towards the further scope of quantum computing, driven quantum algorithms such as quantum neural networks (QNNs) have a great potential for tackling the issue of the classification of bona fide or synthetic images, audio, and videos. However, the field is still evolving, and practical quantum computers may face challenges and limitations. Algorithms or models prepared for DeepFakes according to the quantum architecture depend on the specific techniques employed for classifying the pure-set and false-set from the multi-modal corpus. Consequently, classical computers may struggle to detect such DeepFakes if quantum algorithms exploit the unique properties of quantum systems. As the improvement of quantum computing is ongoing, continuous research is needed to understand its full implications in AI, especially vision intelligence. It would be another hot

topic of further research for designing compatible and adaptable algorithms for DeepFakes that will work for both classical and quantum architectures or systems.

Author Contributions: Author Contributions: Conceptualization, G.G. and K.R.; methodology, K.R. and M.G.; software, G.G. and K.R.; validation, M.G. and T.J.; formal analysis, G.G. K.R. and M.G.; investigation, G.G. and M.G.; resources, T.J.; data curation, M.G. and T.J.; writing—original draft preparation, G.G. and K.R.; writing—review and editing, T.J. and S.T.W.; visualization, M.G. and T.J.; supervision, T.J. and S.T.W.; project administration, M.G. and M.P.; funding acquisition, T.J. and M.P. 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. Ajao, O.; Bhowmik, D.; Zargari, S. Sentiment aware fake news detection on online social networks. In Proceedings of the ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 2507–2511.
2. Vaccari, C.; Chadwick, A. Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Soc. Media+ Soc.* **2020**, *6*, 2056305120903408. [CrossRef]
3. Eelmaa, S. Sexualization of Children in Deepfakes and Hentai: Examining Reddit User Views. *SocArxiv* **2021**, *10*.
4. Nguyen, T.T.; Nguyen, Q.V.H.; Nguyen, D.T.; Nguyen, D.T.; Huynh-The, T.; Nahavandi, S.; Nguyen, T.T.; Pham, Q.V.; Nguyen, C.M. Deep learning for deepfakes creation and detection: A survey. *Comput. Vis. Image Underst.* **2022**, *223*, 103525. [CrossRef]
5. Yu, P.; Xia, Z.; Fei, J.; Lu, Y. A survey on deepfake video detection. *Iet Biom.* **2021**, *10*, 607–624. [CrossRef]
6. Brock, A.; Lim, T.; Ritchie, J.M.; Weston, N. Neural photo editing with introspective adversarial networks. *arXiv* **2016**, arXiv:1609.07093.
7. Afzal, S.; Ghani, S.; Hittawe, M.M.; Rashid, S.F.; Knio, O.M.; Hadwiger, M.; Hoteit, I. Visualization and Visual Analytics Approaches for Image and Video Datasets: A Survey. *ACM Trans. Interact. Intell. Syst.* **2023**, *13*, 5. [CrossRef]
8. Akhtar, Z. Deepfakes Generation and Detection: A Short Survey. *J. Imaging* **2023**, *9*, 18. [CrossRef]
9. Narayan, K.; Agarwal, H.; Thakral, K.; Mittal, S.; Vatsa, M.; Singh, R. DF-Platter: Multi-Face Heterogeneous Deepfake Dataset. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 9739–9748.
10. Li, Y.; Chang, M.C.; Lyu, S. In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In Proceedings of the 2018 IEEE International Workshop on Information Forensics and Security (WIFS), Hong Kong, China, 11–13 December 2018; pp. 11–13.
11. Afchar, D.; Nozick, V.; Yamagishi, J.; Echizen, I. Mesonet: A compact facial video forgery detection network. In Proceedings of the 2018 IEEE International Workshop on Information Forensics and Security (WIFS), Hong Kong, China, 11–13 December 2018; pp. 1–7.
12. Hinton, G.E.; Krizhevsky, A.; Wang, S.D. Transforming auto-encoders. In Proceedings of the Artificial Neural Networks and Machine Learning–ICANN 2011: 21st International Conference on Artificial Neural Networks, Espoo, Finland, 14–17 June 2011; Proceedings, Part I 21; Springer: Berlin/Heidelberg, Germany, 2011; pp. 44–51.
13. Nguyen, H.H.; Yamagishi, J.; Echizen, I. Capsule-forensics: Using capsule networks to detect forged images and videos. In Proceedings of the ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 12–17 May 2019; pp. 2307–2311.
14. Rossler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; Nießner, M. Faceforensics++: Learning to detect manipulated facial images. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Repubic of Korea, 27 October–2 November 2019; pp. 1–11.
15. Dolhansky, B.; Howes, R.; Pflaum, B.; Baram, N.; Ferrer, C.C. The deepfake detection challenge (dfdc) preview dataset. *arXiv* **2019**, arXiv:1910.08854.
16. Korshunov, P.; Marcel, S. Vulnerability assessment and detection of deepfake videos. In Proceedings of the 2019 International Conference on Biometrics (ICB), Crete, Greece, 4–7 June 2019; pp. 1–6.
17. Tariq, S.; Lee, S.; Woo, S.S. A convolutional lstm based residual network for deepfake video detection. *arXiv* **2020**, arXiv:2009.07480.
18. Agarwal, S.; Varshney, L.R. Limits of deepfake detection: A robust estimation viewpoint. *arXiv* **2019**, arXiv:1905.03493.
19. Lyu, S. Deepfake detection: Current challenges and next steps. In Proceedings of the 2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), London, UK, 6–10 July 2020; pp. 1–6.
20. Kumar, A.; Bhavsar, A.; Verma, R. Detecting deepfakes with metric learning. In Proceedings of the 2020 8th International Workshop on Biometrics and Forensics (IWBF), Porto, Portugal, 29–30 April 2020; pp. 1–6.

21. Mittal, T.; Bhattacharya, U.; Chandra, R.; Bera, A.; Manocha, D. Emotions don't lie: An audio-visual deepfake detection method using affective cues. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 2823–2832.
22. Kawa, P.; Syga, P. A note on deepfake detection with low-resources. *arXiv* **2020**, arXiv:2006.05183.
23. Chugh, K.; Gupta, P.; Dhall, A.; Subramanian, R. Not made for each other-audio-visual dissonance-based deepfake detection and localization. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 439–447.
24. Kaur, S.; Kumar, P.; Kumaraguru, P. Deepfakes: Temporal sequential analysis to detect face-swapped video clips using convolutional long short-term memory. *J. Electron. Imaging* **2020**, *29*, 033013. [CrossRef]
25. Symeon, P.C.G.K.Z.; Kompatsiaris, P.I. AFace PREPROCESSING APPROACH FOR IMPROVED DEEPFAKE DETECTION. *arXiv* **2020**, arXiv:2006.07084.
26. Rahul, U.; Ragul, M.; Vignesh, K.; Tejeswini, K. Deepfake video forensics based on transfer learning. *Int. J. Recent Technol. Eng. (IJRTE)* **2020**, *8*, 5069–5073.
27. Wubet, W.M. The deepfake challenges and deepfake video detection. *Int. J. Innov. Technol. Explor. Eng.* **2020**, *9*, 789–796. [CrossRef]
28. Pishori, A.; Rollins, B.; van Houten, N.; Chatwani, N.; Uraimov, O. Detecting deepfake videos: An analysis of three techniques. *arXiv* **2020**, arXiv:2007.08517.
29. Hussain, S.; Neekhara, P.; Jere, M.; Koushanfar, F.; McAuley, J. Adversarial deepfakes: Evaluating vulnerability of deepfake detectors to adversarial examples. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Virtual Conference, 5–9 January 2021; pp. 3348–3357.
30. Vakhshiteh, F.; Ramachandra, R.; Nickabadi, A. Threat of adversarial attacks on face recognition: A comprehensive survey. *arXiv* **2020**, arXiv:2007.11709.
31. Mirsky, Y.; Lee, W. The creation and detection of deepfakes: A survey. *ACM Comput. Surv. (CSUR)* **2021**, *54*, 1–41. [CrossRef]
32. Tolosana, R.; Vera-Rodriguez, R.; Fierrez, J.; Morales, A.; Ortega-Garcia, J. Deepfakes and beyond: A survey of face manipulation and fake detection. *Inf. Fusion* **2020**, *64*, 131–148. [CrossRef]
33. Sohrawardi, S.J.; Seng, S.; Chintha, A.; Thai, B.; Hickerson, A.; Ptucha, R.; Wright, M. Defaking DeepFakes: Understanding journalists' needs for DeepFake detection. In Proceedings of the Computation+ Journalism 2020 Conference, Boston, MA, USA, 20–21 March 2020; Volume 21.
34. Verdoliva, L. Media forensics and deepfakes: An overview. *IEEE J. Sel. Top. Signal Process.* **2020**, *14*, 910–932. [CrossRef]
35. Neves, J.C.; Tolosana, R.; Vera-Rodriguez, R.; Lopes, V.; Proença, H.; Fierrez, J. Ganprintr: Improved fakes and evaluation of the state of the art in face manipulation detection. *IEEE J. Sel. Top. Signal Process.* **2020**, *14*, 1038–1048. [CrossRef]
36. Güera, D.; Delp, E.J. Deepfake video detection using recurrent neural networks. In Proceedings of the 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Auckland, New Zealand, 27–30 November 2018; pp. 1–6.
37. Guarnera, L.; Giudice, O.; Battiato, S. Deepfake detection by analyzing convolutional traces. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 666–667.
38. Huang, Y.; Juefei-Xu, F.; Wang, R.; Guo, Q.; Ma, L.; Xie, X.; Li, J.; Miao, W.; Liu, Y.; Pu, G. Fakepolisher: Making deepfakes more detection-evasive by shallow reconstruction. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 1217–1226.
39. Masi, I.; Killekar, A.; Mascarenhas, R.M.; Gurudatt, S.P.; AbdAlmageed, W. Two-branch recurrent network for isolating deepfakes in videos. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part VII 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 667–684.
40. Trinh, L.; Tsang, M.; Rambhatla, S.; Liu, Y. Interpretable and trustworthy deepfake detection via dynamic prototypes. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Virtual Conference, 5–9 January 2021; pp. 1973–1983.
41. Li, L.; Bao, J.; Zhang, T.; Yang, H.; Chen, D.; Wen, F.; Guo, B. Face X-ray for more general face forgery detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 5001–5010.
42. Li, L.; Bao, J.; Yang, H.; Chen, D.; Wen, F. Advancing high fidelity identity swapping for forgery detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 5074–5083.
43. Yang, Z.; Liang, J.; Xu, Y.; Zhang, X.Y.; He, R. Masked relation learning for deepfake detection. *IEEE Trans. Inf. Forensics Secur.* **2023**, *18*, 1696–1708. [CrossRef]
44. 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]
45. Wang, T.; Cheng, H.; Chow, K.P.; Nie, L. Deep convolutional pooling transformer for deepfake detection. *ACM Trans. Multimed. Comput. Commun. Appl.* **2023**, *19*, 179. [CrossRef]
46. 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**, *25*, 8487–8498. [CrossRef]
47. Yu, Y.; Liu, X.; Ni, R.; Yang, S.; Zhao, Y.; Kot, A.C. PVASS-MDD: Predictive Visual-audio Alignment Self-supervision for Multimodal Deepfake Detection. *IEEE Trans. Circuits Syst. Video Technol.* **2023**. [CrossRef]
48. Stupp, C. Fraudsters used AI to mimic CEO's voice in unusual cybercrime case. *Wall Str. J.* **2019**, *30*.

49. Chettri, B.; Mishra, S.; Sturm, B.L.; Benetos, E. A study on convolutional neural network based end-to-end replay anti-spoofing. *arXiv* **2018**, arXiv:1805.09164.
50. Tom, F.; Jain, M.; Dey, P. End-To-End Audio Replay Attack Detection Using Deep Convolutional Networks with Attention. In Proceedings of the Interspeech, Hyderabad, India, 2–6 September 2018; pp. 681–685.
51. Alzantot, M.; Wang, Z.; Srivastava, M.B. Deep residual neural networks for audio spoofing detection. *arXiv* **2019**, arXiv:1907.00501.
52. Todisco, M.; Wang, X.; Vestman, V.; Sahidullah, M.; Delgado, H.; Nautsch, A.; Yamagishi, J.; Evans, N.; Kinnunen, T.; Lee, K.A. ASVspoof 2019: Future horizons in spoofed and fake audio detection. *arXiv* **2019**, arXiv:1904.05441.
53. Balamurali, B.; Lin, K.E.; Lui, S.; Chen, J.M.; Herremans, D. Toward robust audio spoofing detection: A detailed comparison of traditional and learned features. *IEEE Access* **2019**, *7*, 84229–84241. [CrossRef]
54. Kamble, M.R.; Pulikonda, A.K.S.; Krishna, M.V.S.; Patil, H.A. Analysis of Teager Energy Profiles for Spoof Speech Detection. In Proceedings of the Odyssey, Tokyo, Japan, 1–5 November 2020; pp. 304–311.
55. Chen, T.; Kumar, A.; Nagarsheth, P.; Sivaraman, G.; Khoury, E. Generalization of Audio Deepfake Detection. In Proceedings of the Odyssey, Tokyo, Japan, 1–5 November 2020; pp. 132–137.
56. Rupesh Kumar, S.; Bharathi, B. A novel approach towards generalization of countermeasure for spoofing attack on asv systems. *Circuits Syst. Signal Process.* **2020**, *40*, 872–889.
57. Chintha, A.; Thai, B.; Sohrawardi, S.J.; Bhatt, K.; Hickerson, A.; Wright, M.; Ptucha, R. Recurrent convolutional structures for audio spoof and video deepfake detection. *IEEE J. Sel. Top. Signal Process.* **2020**, *14*, 1024–1037. [CrossRef]
58. Das, R.K.; Yang, J.; Li, H. Long range acoustic and deep features perspective on ASVspoof 2019. In Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Singapore, 14–18 December 2019; pp. 1018–1025.
59. Li, Y.; Yang, X.; Sun, P.; Qi, H.; Lyu, S. Celeb-df: A large-scale challenging dataset for deepfake forensics. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 3207–3216.
60. Cozzolino, D.; Rössler, A.; Thies, J.; Nießner, M.; Verdoliva, L. Id-reveal: Identity-aware deepfake video detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 15108–15117.
61. Zhao, H.; Zhou, W.; Chen, D.; Wei, T.; Zhang, W.; Yu, N. Multi-attentional deepfake detection. In Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition, Montreal, QC, Canada, 10–17 October 2021; pp. 2185–2194.
62. Zhou, Y.; Lim, S.N. Joint audio-visual deepfake detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 14800–14809.
63. Zhao, L.; Zhang, M.; Ding, H.; Cui, X. MFF-Net: Deepfake detection network based on multi-feature fusion. *Entropy* **2021**, *23*, 1692. [CrossRef] [PubMed]
64. Nadimpalli, A.V.; Rattani, A. On improving cross-dataset generalization of deepfake detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 91–99.
65. Ju, Y.; Jia, S.; Cai, J.; Guan, H.; Lyu, S. GLFF: Global and Local Feature Fusion for Face Forgery Detection. *arXiv* **2022**, arXiv:2211.08615.
66. Yan, Z.; Zhang, Y.; Fan, Y.; Wu, B. UCF: Uncovering Common Features for Generalizable Deepfake Detection. *arXiv* **2023**, arXiv:2304.13949.
67. Mcuba, M.; Singh, A.; Ikuesan, R.A.; Venter, H. The Effect of Deep Learning Methods on Deepfake Audio Detection for Digital Investigation. *Procedia Comput. Sci.* **2023**, *219*, 211–219. [CrossRef]
68. Porcu, S.; Floris, A.; Atzori, L. Evaluation of data augmentation techniques for facial expression recognition systems. *Electronics* **2020**, *9*, 1892. [CrossRef]
69. Ilahi, I.; Usama, M.; Qadir, J.; Janjua, M.U.; Al-Fuqaha, A.; Hoang, D.T.; Niyato, D. Challenges and countermeasures for adversarial attacks on deep reinforcement learning. *IEEE Trans. Artif. Intell.* **2021**, *3*, 90–109. [CrossRef]
70. Wang, L.; Meng, X.; Li, D.; Zhang, X.; Ji, S.; Guo, S. DEEPFAKER: A Unified Evaluation Platform for Facial Deepfake and Detection Models. *ACM Trans. Priv. Secur.* **2023**. [CrossRef]
71. Muthukkumarasamy, V.; Sudarsan, S.D.; Shyamasundar, R.K. *Information Systems Security: 19th International Conference, ICISS 2023, Raipur, India, December 16–20. 2023, Proceedings*; Springer Nature: Berlin/Heidelberg, Germany, 2023; Volume 14424.
72. Zhang, L.; Qiao, T.; Xu, M.; Zheng, N.; Xie, S. Unsupervised learning-based framework for deepfake video detection. *IEEE Trans. Multimed.* **2022**, *25*, 4785–4799. [CrossRef]