International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

A Comprehensive Survey on Deepfake Detection Techniques

Duha A. Sultan[1], Prof. Dr. Laheeb M. Ibrahim[2]

Submitted: 12/09/2022 **Accepted**: 16/12/2022

Abstract:

Improving machine learning and artificial intelligence makes it possible to swap someone else's face and voice in a high realism video which made distinguishing the difference between the real and fake videos difficult. Although this technology can be used in many useful fields like advertising, video gaming, and film industry, most of the time it is used for malicious purposes. Therefore, many studies have been done to understand how deepfake works and how to detect these fake videos or images. In this paper, an inclusive study is presented on the existing techniques used for creating and detecting fake materials and analyzing these techniques that are used by several researchers in addition to the great role of artificial intelligence and deep learning on improving them.

Keywords: Machine learning, Deep learning, Deepfake, GAN, Deepfake detection, Forgery detection, Convolutional Neural Network.

1. Introduction:

Deepfake is the digital media, including photos or videos that are manipulated in such a way where a person's likeness substitutes that of another. This manipulation is done by artificial intelligence techniques known as deep learning.

Lip_syncing, Face swapping and head puppetry are the three major kinds of deepfake videos [1]. Different architectures were used to create deepfakes, the most popular ones are:

1- Autoencoder-decoder**,** during training, similar encoder-decoder pair (for both persons A and B) is utilized for learning the latent features of faces, although during generation, the decoders are switched, the latent face A is subjected to decoder B for generating face A similar to face B[2].

2- The second method that produces fake video at a high degree of perfection is *Generative Adversarial Networks* (GAN). It is a system with two neural networks (Generator and Discriminator) sparred against each other like a game contest. The generator learns making the target output when the discriminator distinguishes true from the generator outputs [3]. Goodfellow et. Al first introduced GAN in 2014.[4], and opened up a new research field. In few years, plenty of papers come up on this topic and generative adversarial models showed promising results in the generation of realistic images and videos. Table1 shows some of the most popular GANs.

[1]Dept. of Biology, Education College for Girls University of Mosul, Mosul, Iraq
duhaasultan@uomosul.edu.iq
[2]Dept. of Software Engineering, College of Computer Sciences and Mathematics, University of Mosul, Mosul, Iraq
Laheeb_alzubaidy321966@uomosul.edu.iq

Table 1. Common GANs

GAN type	Description	Year	Function
FCGAN [4]	Fully Connected GAN (traditional GAN)	2014	The generator and discriminator fully connected neural networks are utilized. Used with simple datasets, like MNIST
cGAN [5]	Conditional GAN	2014	Using conditioning on the generator and the discriminator by supplying each with class labels.
DCGAN[6]	Deep Convolutional GAN	2015	First works performing a deconvolutional neural networks, enabling the image creation with higher resolution.
BiGAN [7]	Bidirectional GAN	2016	Allow the inverse mapping, i.e., Data is projected back into the latent space.
SGAN [8]	Semi-supervised GAN	2016	Suggested in the case of semi-supervised learning Discriminator is multiheaded, ie.,has softmax and sigmoid for classification.
WGAN [9]	Wasserstein GAN	2017	Improves the stability of GAN training
WGAN-GP [10]	Wasserstein GAN-Gradient Penalty	2017	Better stability of GAN training than WGAN.
LSGAN [11]	Least Square GAN	2017	Remedying the vanishing gradient problems (use least square loss than sigmoid cross entropy loss)
PGGAN [12]	Progressive growing GAN	2017	Growing the network progressively Improving the quality, stability, and variation
StyleGAN[13]	Style-based generator architecture for GAN	2018	Introduced by NVIDIA research. Using ProGAN plus image style transfer and adaptive instance normalization(AdaIn) . controlling over the generated image style
SAGAN [14]	Sef-Attention GAN	2018	Self-attention mechanism in designing the generator and the discriminator. able to learn global, long-range dependencies for images generation. Attaining major performances on the generation of the multi-class images.
BigGAN [15]	Large-scale GAN	2018	Large-scale GAN training for synthesizing natural images with high fidelity. Create high fidelity and high resolution images
StarGAN[16]	Unified GAN	2018	Allowing concurrent training of many datasets with various domains in one single network.
AttGAN [17]	Attribute GAN	2018	Facial attribute editing manipulates single or multiple attributes, i.e., generating new faces with desired attributes during the preservation of other details.
StyleGAN2 [18]	Improvement of StyleGAN	2019	Enhances the original StyleGAN by many improvements of the aspects: regularization, normalization, and progressively growing techniques.
GDWCT [19]	Group-wise Deep Whitening-and-Coloring Technique	2019	End-to-end training in image translations to convey the profound style semantics. Simple forward propagation- high image quality- memory time efficiency.
MSG-GAN [20]	Multi-Scale Gradient for GAN	2020	Improves training stability Generates high resolution images on many image datasets of various sizes, resolutions, and domains.

Many applications in the field of deepfake creation appeared and spread through the internet. The first one was FakeApp allowing users swap faces. More similar applications over time created including FaceSwap, DeepFaceLab, DFaker, and many more making so easy for everyone to create fake videos (even if s(he) has no knowledge on this subject). This led to the need to find ways that help differentiate between the fake video from the real ones. Recently, many researchers have done all their efforts in this field. This work dealt with the problem of deepfake detection using different techniques: like machine learning techniques based on SVM, deeplearning techniques like Convolutional Neural Network (CNN), CNN with SVM, CNN with LSTM, and RNN, under two groups: image and video detection methods. A brief explanation on the most common dataset is presented which will serve as an important reference for those interested in this field.

The paper consists of: section 1 is an introduction. Section 2 details deepfake detection methods with their related works. Common datasets are in section 3. Finally,

the conclusions and our future expectations are in section 4.

2. Deepfake Detection

Creating deep fake videos is simple, although in terms of detection, it's a key challenge. Early attempts to detect fake images or videos depended on handcrafted features gained from inconsistencies and artifacts during synthesis of the fake videos. In the modern methods, deep learning automatically extracted discriminative and salient features for detecting deepfakes. To train classification models, this technique required large dataset of fake and real videos. Regardless of the technology used, deepfake detection methods can be: *fake image* and *fake video* detection.

2.1. Fake image Detection

To detect fake images, different methods have been proposed, especially those generated by GAN. Most of detecting methods were based on deep learning techniques for extracting face topographies within the image and try to determine the inconsistencies within it.

Guarnera L. et. al.[21], suggest a detection method by means of Expectation Maximization(EM) algorithm [22], extracting local landscapes to model convolutional generative traces in GAN-generated images. Six datasets for training and testing, CelebA[23], for authentic face images while fake images were generated using five different GAN-architectures: STYLEGAN[13], STYLEGAN2[18], STARGAN[16], ATTGAN[17] and GDWCT[19]. The researchers compare among these datasets using three types of classifiers (K-NN, SVM, LDA) as in Fig.1.

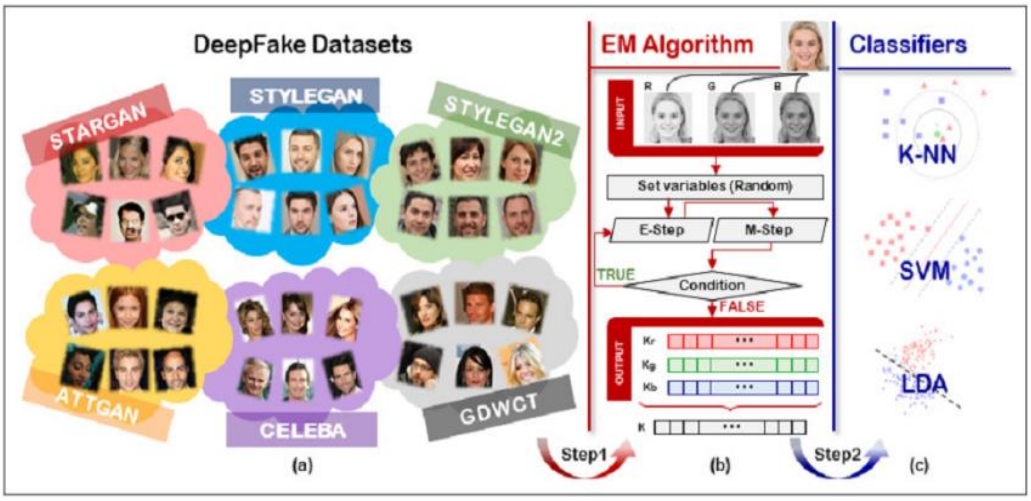

Figure 1. General pipeline. (a) Real Datasets (CELEBA) and Deepfake images, (b) Every images in (a) topographies extracted by EM algorithm; (c) classifiers (K-NN, SVM, LDA).

Hsu C.C. et. al.[24], submitted a detection method according to deep learning technique by integrated Siamese networks [25] with the DenseNet[26] architecture and a contrastive losses for learning the network by pairwise learning strategies (Developed DenseNet was used to allow pairwise information as input). Several GANs are used to create the real-fake image pairs. The extracted features are then fed to a classifier (Fig.2) concatenated to the last network layer for revealing whether the input image is real or fake. The used dataset was extracted from CelebA, while five GANs were used to generate fake images: PGGAN[12], LSGAN[11], WGAN[9], WGAN-GP[10], and DCGAN[6], every GAN randomly created 40,000 fake images. As a total 200,000 fake images were used, also another 200,000 real images were selected by the researchers from CelebA. The overall images then split into 380,000 to train, 10,000 to validate, and 10,000 to test with equal numbers of real and fake images in every set. The researchers confirmed designed method to have higher generalization ability and effectiveness than alternative approaches.

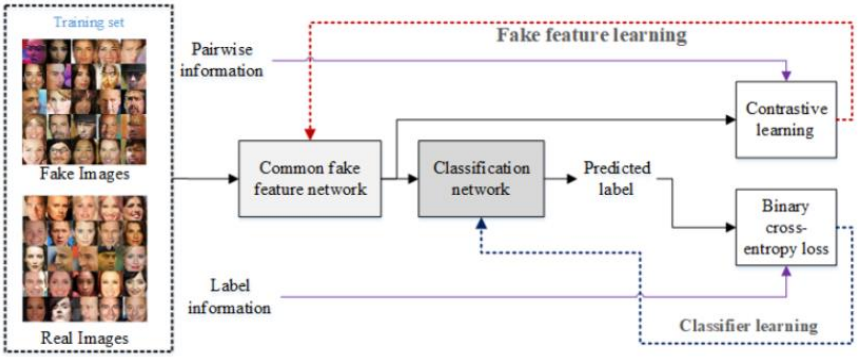

Figure 2. fake detector based on the two-step learning approach with the proposed fake feature network.

Li L. et. al.[27] proposed an alternative approach in detecting forged images named as *face X-ray*, when the image is created by mixing two images, then intrinsic image inconsistencies exist across the image boundaries. So, a forged face image can be detected by discovering the blending boundaries using conflicts of the image statistics across the boundaries, instead of capturing the artifacts that are synthesized through manipulations. Without using the fake pictures, only blended ones are used to train the model. The researchers prove that the method showed high detection accuracies with better generalization ability.

As the evolving of Gan is continuous, new versions appeared and this led to the need to consider and develop the detection models' generalization ability. Xuan et. al. [28] focused in their research on this problem using CNN with a preprocessing step in the training stage by adding Gaussian noise or Gaussian blur for destroying low high frequency artifacts in GAN images. This can improve low level similarity between the fake and real pictures, forcing more essential features to be learned by the classifier that led to better generalization ability. The general framework of the model is illustrated in fig.3 where in the dataset: real images are extracted from CelebA, which includes high quality face images, while the fake images are generated by PGGAN, WGAN-GP and DCGAN.

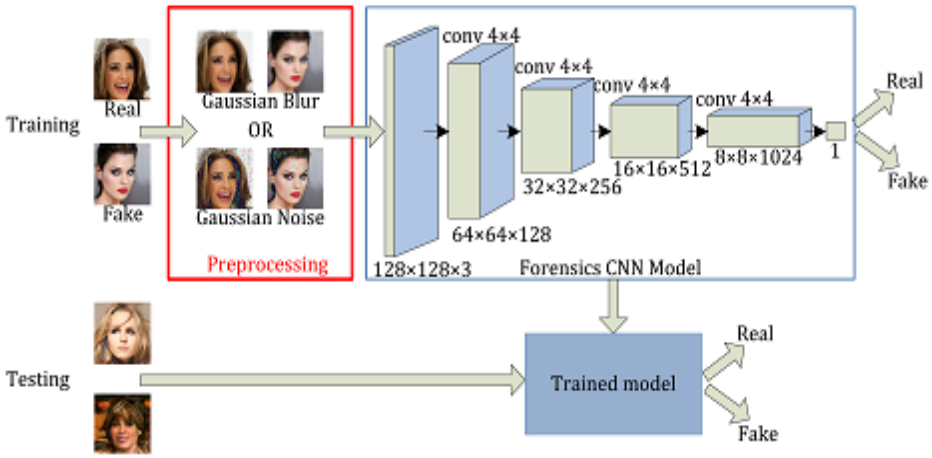

Figure.3. The general framework of the designed method

Zhao et. al.[29] offered a new detection method for deepfake images using the indication on the source trait inconsistency within the fake images. They assumed that these traits are still kept on the modified image after the deepfake generation process. For this, a faked image could include dissimilar source traits at various positions, while an original image is consistent in all positions. So, fake images can be detected by extracting the local source traits and measuring their self-consistency. To extract these traits , a CNN with a pair-wise self-consistency learning (PCL) approach were used combined with a new image synthesis approach named as I2G: inconsistency image generator to supply the training data required for PCL. Three of the most common dataset were used, which are FaceForensics++ (FF++)[30], DFDC[31] and CD2. the researchers indicated the PCL and I2G compete other state-of-the-art approaches providing a baseline for future works.

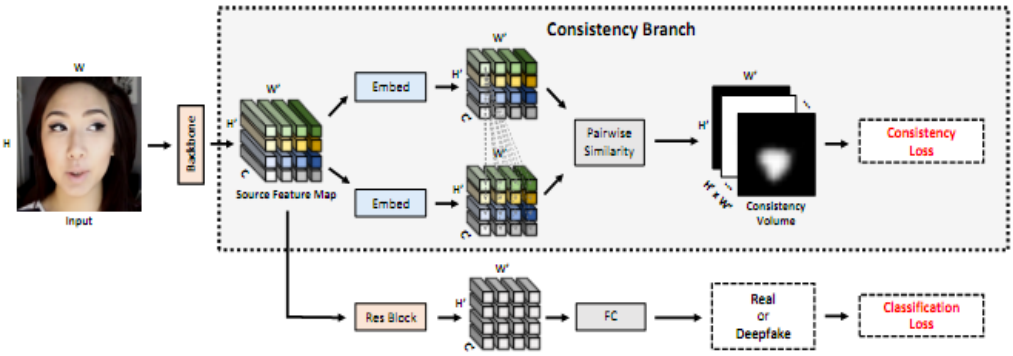

Figure.4. PCL architecture

Previous methods rely for deepfake detection on small traces due to imperfect facial forgery approaches. Alternatives could have no traces with dvarious features in various face image areas. It is not realistic for the collecting of all forgery apprahces as much as possible. To overcome this problem, Sitong L. et. al.[32] offered a random shuffling method as an alternative perspective in deepfake detection. Each image is divided into blocks (196 blocks), then offered a random shuffling to intra_block and inter_block. This model enhances network robustness by intra_block shuffling which healps the network at learning local features by inter_block shuffling as fig.5 expalins. The method is not dependent on specific classification networks. Xception is adopted as a backbone network. Yet the model mainly focued the generalization ability and obtained good performance at 99.72% of AUC on FF++ and Celeb_DF[33] datasets.

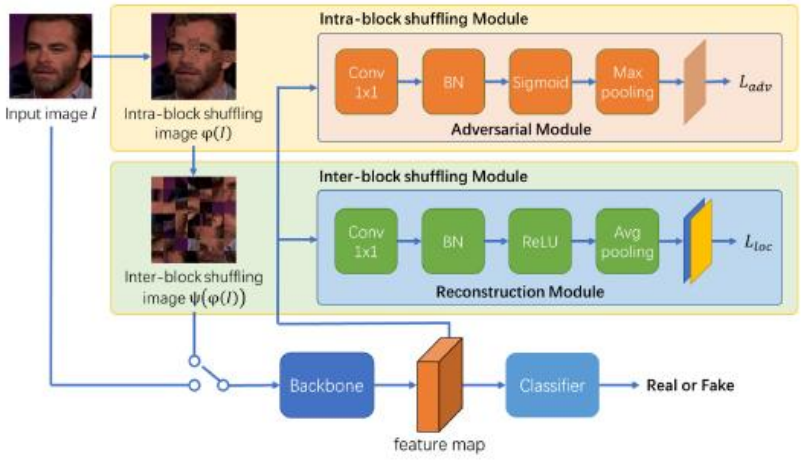

Figure 5. The framework of block shuffling for deepfake detection

2.2. Fake Video Detection

In fake videos there will surely be something off (that is not consistence with the overall video). The goal is to find these inconsistence artifacts. Some significant variations between videos and images make most image detection methods unusable with videos:

1- Videos are heavily compressed than images, hiding traces of forgery.
2- Videos are series of frames (images) with multiple temporal constraints: with important correlation between subsequent images as the individual image quality.

For this, deepfake video detections are divided into: those that use *temporal features* to take advantage of the temporal relationship across multiple frames, and those that utilize *visual artifacts* in the frames.

2.2.1. Visual Artifacts within a frame

Deepfakes usually produces artifacts that are not easy for humans to identify but can be determined quickly by machine and forensic analysis. GAN fingerprints background irregularities, and inconsistent features in the forged face images; like blurred and different skin colors are all examples of visual artifacts. Detecting facial tampering in videos, especially those Deepfake and Face2Face techniques create, Afchar D . et.al.[34] submitted in their research a method built on two convolutional neural network, name as Meso-4 and MesoInception-4. Two datasets were used for deepfake, 175 forged videos collected from various platforms of 2 seconds to 3 minutes and with resolution of at least 840x480 pixels. Real face images also extracted from internet having identical resolution. The second dataset is face forensic which contains over 1000 fake and real videos created by Face2Face approach. Face Viola-Jones[35] detector was used to extract face region and

aligned with facial land mark detection. 50 faces were extracted per scene and a high and successful detection rate were obtained of about 98% for deepfake and 95% for Face2Face.

Another architecture based on CNN introduced by Aya I. et.al.[36] for effective deepfake detection named as YOLO-CNN-XGBoost, as seen in Fig.6. The aim is to integrate the benefits and pros of both XGBoost[37] and CNN models inorder to improve deepfake video detection. YOLO face detector[38] determined faces in video frames, while InceptionResNetV2 model extracted the discriminant spatial-visual features aid to explore the visual artifacts in the video frames, fed to the XGBoost classifier for distinguishing between the real and fake videos. This method showed 90.62% of AUC, 90.73% accuracy, and 86.36% F1-measure on the Celeb-DF, FaceForencics++(c23) combined dataset.

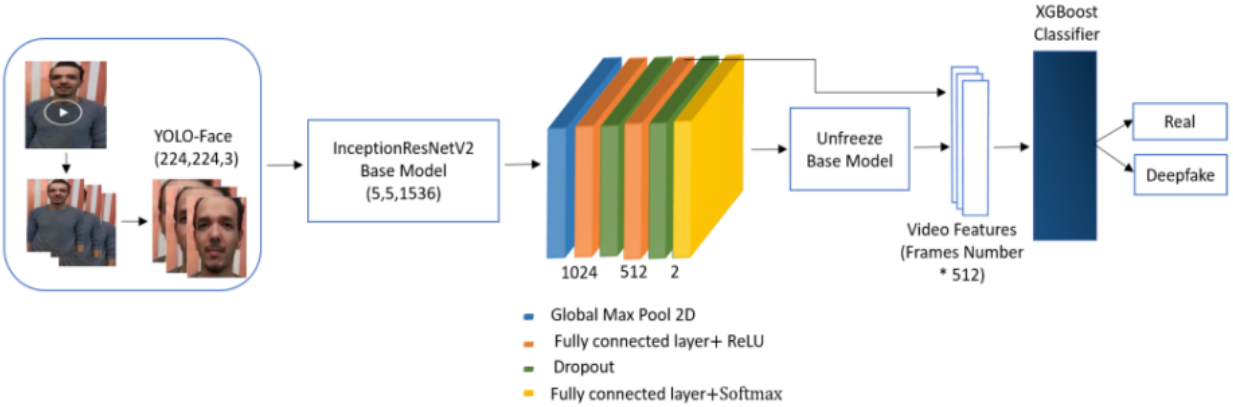

Figure 6. YOLO-CNN-XGBoost Deepfake videos detection system architecture

Amerini et.al.[39], presented a method that uses optical flow vectors calculated for two consecutive frames, Fig.7 illustrate the idea. The authors have a hypothesis that the optical flow is capable of exploiting discrepancies in motion across frames synthetically made in relation to those naturally the video camera creats, the unusual movement of lips, eye, and the whole face. The flow vectors are transformed to color images with three channels using a fixed color coding approach, so that ResNet and VGG16 models extract features. FF++ dataset were used and the model achieved accuracy of 81.61% and 75.46% for both VGG16 and ResNet50 respectively.

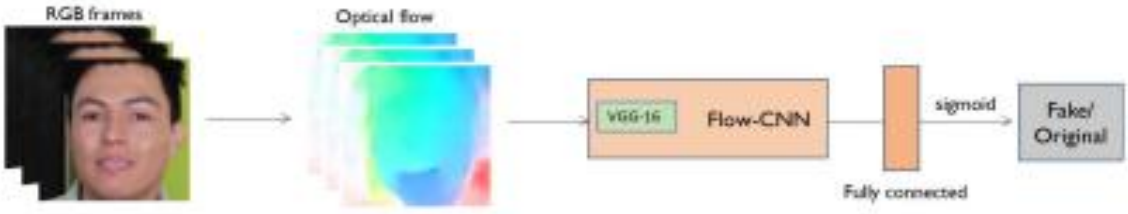

Figure 7. Optical flow-CNN architecture

For enhancing the accuracy of classification and allow the use of a light model, Tran V.N. et. al. [40] proposed an architecture according to the classifier network with manual attention target-specific region for forming distillation set, fig.8. Person detection is implemented by "OpenCV"[41]. The face, within each frame, is then extracted using Multitask cascaded convolutional neural networks (MTCNN)[42]. These two steps will decrease the amount of deceptive data that is used to fool the model. For each face region, facial landmarks are extracted using "Openface2"[43]. Distillation set containing several patches is constructed to determine which parts of the face would be trained, so the input to the classifier are the distillation sets. Inception v3[44] have been used to train the entire face while MobileNet [45] network used to train face patches. The model performance is evaluated on two datasets: small dataset Celeb_DFv2, which includes 590 original videos and 5639 fake videos, and large dataset DFDC available on the internet with over 100,000 videos. The result of testing 0.9628 of F1_score, 0.978 of AUC for Celeb_DF v2, and 0.9243 of F1_score, 0.958 of AUC for DFDC.

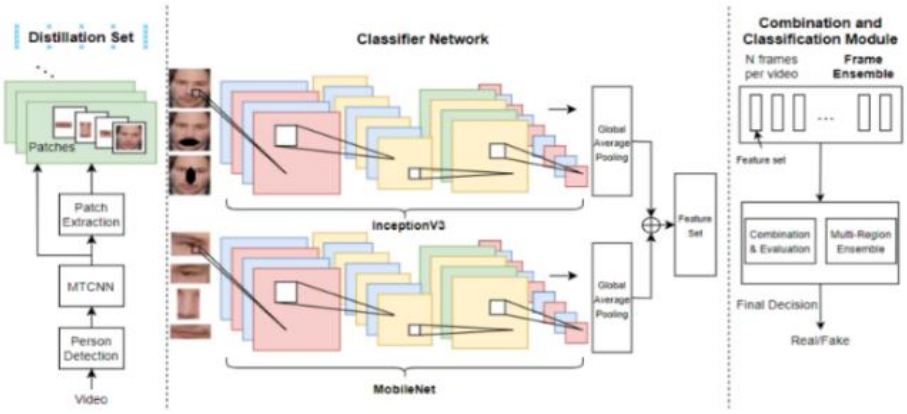

Figure 8. Inception and MobileNet CNN with distillation set deepfake detection model

Matern F.et. al.[46] proposed some straight forward topographies for detecting created faces, Deepfake and Face2Face images. The features that adopted by the researchers are: 1-both eyes are expected to have similar radii, similar color. 2- for the right and left eyes, the iris center of and the eye distance center must be the same. 3- missing reflection and particulars in the eyes and teeth areas. So, Hough circle transform, and Canny edge detection were used to detect eye region before feature extraction. 367 real and 342 fake samples are collected and downloaded from the internet. Neural network and logistic regression were used for classification. The authors proved that, combining features of eyes and teeth yielded better result with an AUC of 0.851 than classifying using the features separately.

Deressa W. et.al.[47] offered a model to detect deepfakes of a Convolutional neural network Vision Transformer (CViT). Learnable features are extracted by the CNN (VGG_like, with no fully connected layer), while the ViT takes , as input, these features and classifies them using an attention. The model is trained on a DFDC dataset and has achieved AUC value of 0.91, 91.5% accuracy, and a loss value of 0.32.

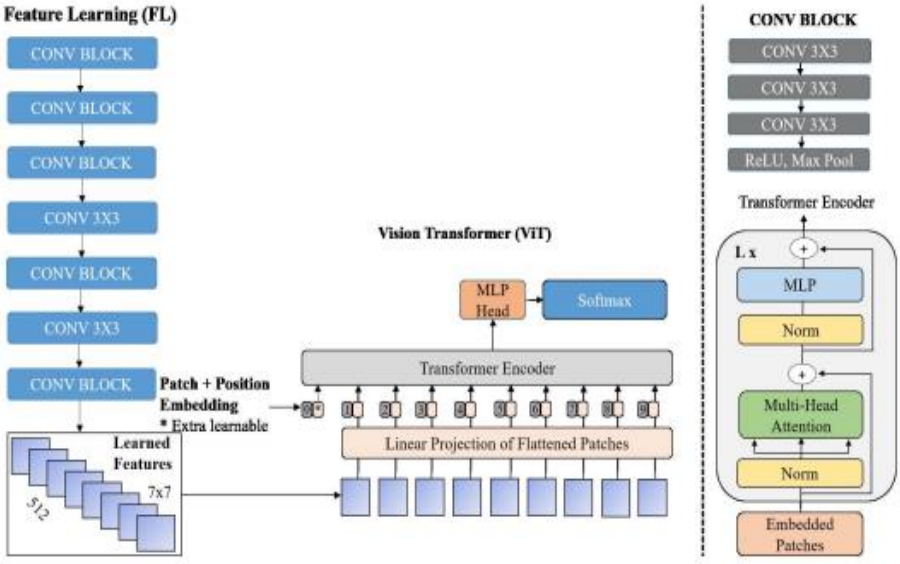

Figure 9. Computer Vision Transformer

2.2.2. Temporal features across multiple frames

Guera D. et. al. [48], presented a two-stage deeplearnig model that combines between CNN and LSTM[49] deep neural network inorder to make use of feature extraction capability of CNN and classification and memorization capability of LSTM recurrent neural network. First features are extracted at a frame level by CNN, then these features are inserted into LSTM network for capturing temporal. Dataset consists of 600 videos, where 300 deepfake videos from multiple websites, incorporated with 300 random real videos from HOHA dataset [50]. The model produced 97% accuracy.

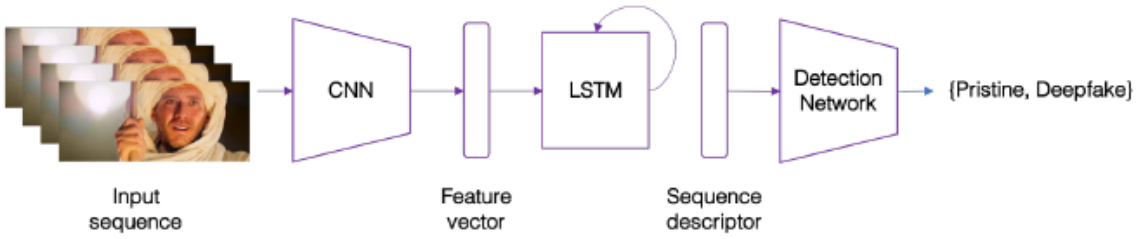

Figure 10. Overview of the detection system

Similar work was done by [51,52]. Abdul Jamsheed V. et. al. [51] used a combined model with ResNeXt[53] and LSTM neural network. worked on a combined dataset taken from DFDC, FaceForensics++, and Celeb_DF datasets. The model succeeded in achieving a result of 92% accuracy. While Priti Y. et.al.[52] proposed a model comprises of InceptionResNet v2 for feature extraction. The output of CNN functions an input to 2048 LSTM layer. The dataset has been collected from DFDC dataset and the model achieved 84.75% and 91.48% accuracy for 20 and 40 epochs respectively.

In Li Y. et al.[54], the CNN_based classifier process is expanded to LRCN[55] by incorporating the temporal relationship between successive frames, as eye blink is temporal since open to close. LRCN is three parts: 1) feature extraction, 2)sequence learning, and 3)state prediction. The eye region is converted into discriminative features via the feature extraction module, which is performed with a VGG16_CNN framework [56]. The output is inserted into sequence learning implemented by RNN with LSTM. The RNN output is then sent to completely linked layers for state prediction. The model was evaluated on CEW dataset [57] which includes 1193 images of closed eyes and 1232 of open eyes. In addition to 50 videos were downloaded containing eye blinking to form EBV dataset. LRCN showed a performance of 0.99.

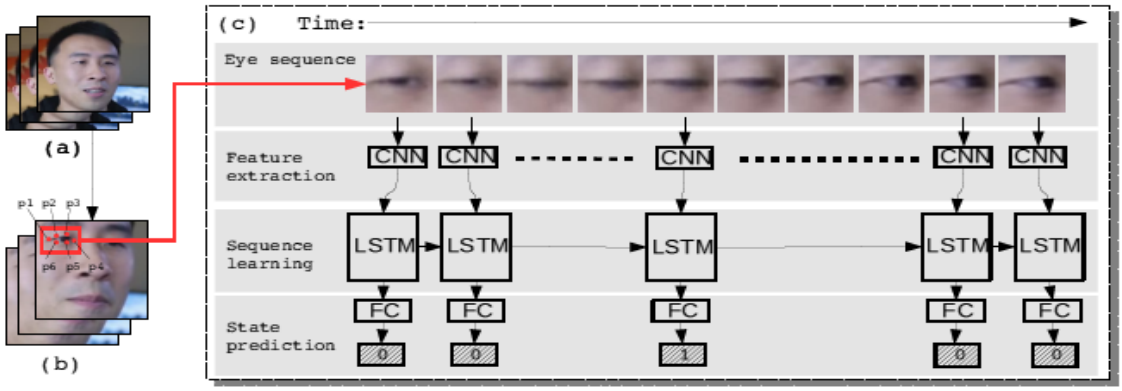

Figure. 11. LRCN method. (a) original sequence (video frames). (b) sequence upon face detection, every frame eye region is determined according to eye landmarks (b) and pass it to (c) LRCN, of three parts: feature extraction, sequence learning and state prediction.

Both methods use CNN to extract features and pass them to RNN layers for temporal processing. Shahrous T. et.al.[58], build a CLRNet model with a Convolutional LSTM according to Residual Network as more comprehensive, as it capture the spatio-temporal information and extract topographies from a sequence of frames. With this technique, there is no need to the use of two different networks (CNN and RNN), as poor results were obtained from these methods when evaluated on unseen deepfake data. To improve generalizability, detection and defense strategies were developed by the authors within learnings of 1)single domain, 2)merge, and 3)transfer. More than 150000 frames from 200 high-quality real-world DFW videos from diverse web sources were used to evaluate the model with a result of 93.22% accuracy.

To make a full use of spatio-temporal information, Temporal Dropout 3-Dimensional Convolution Neural Network (TD-3DCNN) which Daichi Z. et.al.[59] offered. Here, the video frames volumes are sampled by temporal dropout operation and fed into a 3DCNN, which consists of three inception modules, each module of 3x3x3 and 5x5x5 kernel size, for extracting the topographies of various scales and enhancing the modules representation ability. The used dataset were Celeb-DFv2, DFDC, and FF++. The model outperformed six detectors used for comparison, and achieved a competitive performance.

Yipin Z. and Ser-Nam L.[60], achieved more generalized approach, submit a combined visual / auditory deepfake detection task. They observed that there is a strong relationship between the pronounced syllables and the lip motion, when humans speak . At

some moments, this synchronization breaks when any one of the modalities is fake. The proposed framework is to apply central connection to video and audio stream between low level features, which encode spatial (for video frames) and temporal information to higher level semantic representations, i.e. connection of the video and audio networks features representations. Through training, the model automatically learns the correspondence between the audio and the corresponding visual regions. The model is language-agnostic and can be implemented on videos with different languages. No large_scale dataset exists that provides high quality auditory and visual deepfake, so a generated dataset were used by applying various vocoders on the spectrogram of the audio channel of deepfake videos. This was applied on two datasets: FF and DFDC. The model achieved accuracy of 81.96% and 89.55 of AUC.

Other work within this field was submitted by Shruti A. et. al.[61], but the researchers focused of the mouth shape associated with words having the sound M, B, or P in which the mouth must completely close inorder to pronounce these phonemes. The dataset consists of fake videos generated from the real ones using three synthesis techniques: audio-to-video (A2V), text-to-video for short and long utterances, T2V-S and T2V-L respectively. To determine if the expected mouth close is present in any of the video frames, the researchers used three analysis approaches:1- manual (by analyst), 2-profile: a vertical intensity profile is extracted from the middle of the mouth, which is different when the mouth is open or closed. 3- CNN: where Xception architecture was used for classification. Among the results obtained by the researchers, the profile and CNN techniques performance are high on the A2V dataset with accuracy above 96%.

Finally, Wing et.al.[62] presented a comparison of three 3D-CNN models; named as I3D, ResNet3D, and ResNeXt3D to detect fake videos. FF++ dataset was used, where the forged videos were generated using four manipulation schemes: faceswap, deepfake, face2face, and neural textures. The authors investigated three scenarios for training and testing the models: All, Single, and Cross-manipulation techniques. They showed, through the experimental results, that 3D video CNN outperformed other forgery detection algorithms.

Table 2 summarizes all the mentioned methods in the deepfake detection , the used dataset, and the best result obtained by the researchers.

Table 2
Summary of deepfake detection methods

Reference	Method	Dataset	Input	Best Result
Guarnera L. et. al.[21]	Expectation Maximization (EM) algorithm for feature extraction. 3 classifiers: K_NN, SVM, LDA	CelebA for real images Fake images were generated by 5 GAN architectures: STYLEGAN, STYLEGAN2, STARGAN, ATTGAN, GDWCT	image	99.81% accuracy
Hsu C.C. et. al.[24]	Integrated Siamese network with DenseNet CNN	CelebA for real images Fake images were generated by 5 GAN architectures: PGGAN, LSGAN, WGAN, WGAN-GP, DCGAN	image	98.8% precision 94.8% recall
Li L. et. al.[27]	CNN Find the blending boundary between the original and target feces.	FF++, DFD, DFDC, Celeb-DF	Image	99.17% of AUC
Xuan X. et al.[28]	CNN with preprocessing step by adding Gaussian noise or Gaussian blur to destroy low level high frequency artifacts in GAN images	CelebA-HQ for real images Fake images were generated by: DCGAN, WGAN-GP, PGGAN	Images	95.45% accuracy Improve both TPR, TNR
Zhao T. et. al.[29]	CNN with a pairwise self-consistency learning (PCL)	FF++, CD2, DFDC-P I2G to generate fake (inconsistent) images	Images	99.98% of AUC
Sitong L.[32]	Xception CNN with inter_block and intra_block random shuffling	FF++, Celeb-DV v2	images	98.26% accuracy 99.72% of AUC
Afchar D. et. al.[34]	Viola Jones for face detection Two CNN: Meso_4 and MesoInception for feature extraction	Real images were downloaded from the internet. Fake images collected from FF++ dataset and from some different platforms	videos	Detection rate: 98%

Aya I. et. al.[36]	YOLO-CNN-XGBoost model YOLO for face detection Inception ResNet v2 for feature extraction XGBoost for classification	CelebDF FaceForensics++(c23)	videos	90.73% accuracy 90.62% of AUC
Amerini I. et.al. [39]	Optical flow vectors VGG16 and ResNet CNN	FaceForensics++	Videos	81.61% for VGG16 and 75.46% for ResNet50 accuracy
Tran V. N. et al.[40]	CNN with distillation set of face patches Inception v3 used to train the entire face MobileNet used to train face patches	Celeb-DF v2, DFDC	videos	For Celeb_DF v2: 97.8% of AUC 96.28% of F1_score
Matern F. et. al.[46]	Extracting Visual features for both eyes (like eye color, iris radii,….) and missing reflection in teeth and eye regions. Canny edge detection and Hough circle transform for region detection. Neural network and logistic regression for classification.	Real and fake videos are downloaded from the internet	videos	85.1% of AUC
Derresa W. et. al.[47]	Convolutional neural network Vision Transformer CViT VGG_like Net for feature extraction ViT for categorization	DFDC	videos	91.5% accuracy 91% of AUC Loss value of 0.32
Guera D. et. al.[48]	Two-stage deep learning model: CNN+LSTM CNN for frame level features extraction. LSTM for temporal features of multiple frames	300 fake videos from multiple websites. 300 real videos selected from HOHA dataset.	videos	97% accuracy
Abdul Jamsheed V. et. al.[51]	Combined model with ResNeXt for spatial features and LSTM for temporal features.	DFDC , FaceForensics++, Celeb_DF	videos	92% accuracy
Priti Y. et.al. [52]	Combined model with InceptionResNetv2 for spatial features and LSTM for temporal features.	DFDC	Videos	84.75% & 91.48% accuracy for 20 & 40 epochs
Li Y. et. al.[54]	Eye blinking feature is used to detect fake videos by LRCN: Feature extraction by VGG16 CNN Sequence learning by LSTM State prediction by FC layer	EBV dataset formed from: CEW dataset of closed and open images 50 videos containing eye blinking downloaded from the internet	videos	99% accuracy
Shahrous T. et.al. [58]	CLRNet: Convolutional LSTM cells and Residual blocks Network Transfer learning	Training datasets: DF, DFD, F2F, NT. Evaluation dataset: real-world DFW	Videos	93.22% accuracy
Daichi Z. et.al [59]	Temporal dropout 3D CNN consists of three inception modules	Celeb-DFv2, DFDC, FF++	Video frames volume	Outperformed six detectors
Yipin Z. et. al.[60]	Joint visual/auditory deepfake detection Synchronization between lip motion and the pronounced syllables.	Generated dataset by applying various vocoders on the audio channel spectrogram of deepfake videos. This applied on FF and DFDC	Audio / videos	81.96% accuracy 89.55% of AUC
Shruti A. et. al.[61]	Synchronization between lip motion and the pronounced words having the sounds M,B,	Real videos downloaded from the internet Fake videos generated from the real	Audio / videos	96% accuracy

	or P. Three analysis approaches: manual – profile – Xception CNN	ones using three techniques A2V , T2V-S , T2V-L		
Wang Y. and Dantcheva A. [62]	I3D, ResNet3D, and ResNeXt3D	FF++	Video frames volume	Outperformed other detectors

3. Common Datasets

This section shows the most recent and widely used datasets for deepfake detection that generated using deep learning techniques. Table3 summarizes the most common datasets used by deepfake detection models.

- **CelebA dataset:** Is a big scale dataset [23] of 200k celebrity images, each having 40 attribute annotations. Its images cover large pose alterations and background clutters. CelebA shows big varieties, large quantities, and rich annotations,: 10177 identities, 202599 face images, 5 landmark locations, 40 binary attribute annotations in every image.
- **FaceForensics dataset (FF) :** FaceForensics dataset [63] includes half-million manipulated images from 1, 004 videos. It has two subsets made by Face2Face reenactment approach, namely Source-to-Target Reenactment Dataset, performing the reenactment between two randomly selected videos and the second subset represent the Self Reenactment Dataset using the same video as the sourc video and a target. The whole dataset is 1, 408 videos to train, 300 to validate, and 300 for testing, making 732, 391, 151, 835, and 156, 307 images, in respect.
- **Faceforensics++ (FF++) dataset:** FaceForensics++ [30] extends the FaceForensics dataset and is a public dataset as benchmark for the detection of the realistic fake face images. The set is made of 1, 000 thoroughly chosen videos, most which are YouTube, of about 60% of people are male and the remaining 40% are female. The resolution, is about 55% with 854 × 480, i.e., Video Graphics Array (VGA) resolution, 32, 5% in 1, 280 × 720, i.e., high definition (HD), and 12, 5% in 1, 920 × 1, 080 (fullHD) of resolutions.
- **UADFV dataset:**The UADFV [64] is a synthetic dataset from the University of Albany to aid the detection of fake face videos using physiological cues such as eye blinking. The dataset consists of 49 false videos created with the FakeApp application. Each sequence has a resolution of 294 x 500 pixels and an average of 11.14 seconds.
- **Deepfake-TIMIT dataset:** the Deepfake-TIMIT [65] consists of is videos of: 1)Low quality, which composed of 320 videos with around 200 frames of size 64x64 pixels, and 2)High quality, comprises 320 image sequences of approximately 400 frames of 128x128 pixels.
- **HOHA-based dataset :**Contains 8 classes of human actions from 32 Hollywood movies. Guera and Delp [50] presented a dataset of random 300 videos from the HOHA dataset [52]. It is realistic collection of 16 samples from well-known movies stressing human actions, and 300 deepfake videos from different video websites, totalling 600 videos, with approximately 24 frames per second of 360 × 240 format.
- **Deepfake Detection Challenge (DFDC) dataset:** Facebook's Deepfake Detection Challenge (DFDC) [31] dataset comprises 5000 face manipulated videos for actors . The dataset consists of 66 actors selected depending on the characteristics : 74% female and 26% male and, 68% Caucasians, 20% African-American, 9% west-Asian, and 3% south-Asian. The manipulation was carried out using two face swap approaches: First, makes high quality images with faces near the camera, keeping the source proportion swapping the same faces. Second, produces lower swap quality images. So, the dataset consists of 780 clips for testing, and 4,464 for training purposes with different resolutions and each with 15 seconds length.
- **Celeb-DF dataset:**Celeb-DF [33] is a large-scale, challenging deepfake video dataset created utilizing an enhanced synthesis over celebrities' YouTube videos . The dataset collection includes 5639 high-quality videos over two million frames 256×256 pixels every from 59 celebrities, for various ethnicities and ages for females and males. Each video has a frame 30 frames a second and a total length of about 13 seconds describing different aspects such as orientations, face sizes (in pixels), lighting conditions, and backgrounds.
- **DeeperForensics-1.0 dataset:** DeeperForensics-1.0 [66] is large-scale, rich-diversity, and high-quality dataset for forgery detections. It contains 60, 000 videos and 17.6 million frames of automatic swapped face creations, at resolution of 1, 920 × 1, 080 pixels. The original videos were from 100 actors from 26 countries, of females and males of various skin tones and age of 20 to 45 years. In addition, the made eight naturally expressed feelings, i.e., anger, fear, happiness, disgust, surprise, contempt, sadness, and neutral, in different angles ranging from -90^{o} to $+90^{o}$. Also, specific poses, expressions, and lighting conditions of the source images played a big role in the quality of the dataset.
- **CEW dataset:**Closed Eyes in the Wild Dataset (CEW) [57] is of 1192 subjects when both eyes closed and 1231 with opened eyes. Some challenges are amateur picture taking, occlusions, problems with lighting, posture, and motion blurring.

Table 3. Common dataset

dataset	year	Description
CelebA [23]	2018	Contains 202,599 face images of the size 178x218 from 10,177 celebrities
FaceForensics [63]	2018	Comprises 1,408 videos (732,391 images) for training, 300 videos(151,835 images) for validation, 300 videos (156,307 images) for testing
FaceForensics++ [30]	2018	Comprises 1000 videos selected from YouTube, 60% for male and 40% for female
UADFV [64]	2018	Consists of 49 fake videos created by FakeApp application. Each sequence comprises a resolution of 294x500 and 11.14 seconds on average
Deepfake-TIMIT [65]	2018	Consists of 640 fake videos: 320 low quality videos with 200 frames of 64x64 pixels, and 320 high quality videos with 400 frames of size 128x128 pixels
HOHA [50]	2018	Contains 8 classes of human actions from 32 Hollywood movies
HOHA-based [48]	2018	Consists of 300 videos selected from HOHA, and 300 fake videos selected from multiple websites
DFDC [31]	2019	Consists of 5000 videos, 26% for male and 74% for female for different genders
Celeb-DF [33]	2019	Comprises 5639 high quality videos with 2 million frames of size 256x256 pixels
Deeper-Forensics 1.0 [66]	2020	Comprises 60,000 videos with 1920x1080 resolution, and 17.6 million frames
CEW [57]	2020	Contains 1192 subjects of closed eyes, 1231 of open eyes

4. Conclusion

With the increase in the number of fake videos and the harm they have caused to many people and with people losing trust in videos spread on the Internet, it has become necessary to develop special methods to detect these videos. Generating a fake video is rather easy, but the process of detecting a fake video is a challenge due to the continuous development of fake video production technologies. But no matter how precise these techniques are, they do not reach perfect (as long as the video is fabricated, there is something inconsistent in it). By searching for lack of consistency in the video, the fake video is revealed. The findings showed the use of deep learning methods depend on the use of CNN to extract visual artifacts in a frame and LSTM to extract temporal features across multiple frames gave the best results for the classification of real or fake videos.

References:

[1]. SiweiLyu,"DEEPFAKE DETECTION:CURRENT CHALLENGES AND NEXT STEPS",2020 IEEE International Conference on Multimedia& Expo Workshops(ICMEW).

[2]. Bahar U. M., and Afsana S., " Deep Insight of DeepfakeTechnoloy: A Review", 2020, DUJASE Vol.5 (1&2)13-23.

[3]. Nobert Y., "DeepFake Technology: Complete Guide to Deepfakes, Politics and Social Media", July 6 2019, Computers & Technology (Book).

[4]. Goodfellow I., Pouget_Abadie J., Mirza M., Xu B., Warde_Farly D., Ozair S., CourvilleA., and Bengio Y., "Generative Adversarial Mets", 2014, In Advances in Neural Information Processing Systems, 2672-2680.

[5]. Mehdi Mirza and Simon Osindero. 2014. Conditional Generative Adversarial Nets. arXiv:1411.1784. Retrieved from https://arxiv.org/abs/1411.1784.

[6]. Radford, A., Metz, L., and Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434.

[7]. JeX D., Philipp K., and Trevor D. (2016). Adversarial Feature Learning. arXiv:1605.09782. Retrieved from https://arxiv.org/abs/1605.0978.

[8]. Augustus O. (2016) . Semi-supervised Learning with Generative Adversarial Networks. arXiv:1606.01583. Retrieved from https://arxiv.org/abs/1606.01583.

[9]. Arjovsky, M., Chintala, S., and Bottou, L. (2017, July).Wasserstein generative adversarial networks. In InternationalConference on Machine Learning (pp. 214-223).

[10]. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. (2017). Improved training of Wasserstein GANs. In Advances in Neural Information Processing Systems(pp. 5767-5777).

[11]. Mao, X., Li, Q., Xie, H., Lau, R. Y., Wang, Z., and Paul Smolley, S. (2017). Least squares generative adversarial networks. In Proceedings of the IEEE International Conferenceon Computer Vision (pp. 2794-2802).

[12]. Karras, T., Aila, T., Laine, S., and Lehtinen, J. (2017). Progressive growing of GANs for improved quality, stability, andvariation. arXiv preprint arXiv:1710.10196.

[13]. Karras, T., Laine, S., and Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4401-4410).

[14]. Han Z., Ian G., Dimitris M., and Augustus O. (2018). Self-attention Generative Adversarial Networks. arXiv:1805.08318. Retrieved from https://arxiv.org/abs/1805.08318.

[15]. Andrew B., JeX D., and Karen S. (2018). Large Scale GAN Training for High Fidelity Natural Image Synthesis. arXiv:1809.11096. Retrieved from https://arxiv.org/abs/1809.11096.

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

translation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 8789-8797).

[17]. He, Z., Zuo, W., Kan, M., Shan, S., and Chen, X. (2019). AttGAN: Facial attribute editing by only changing what you want. IEEE Transactions on Image Processing, 28(11), 5464-5478.

[18]. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. (2020). Analyzing and improving the image quality of StyleGAN. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 8110-8119).

[19]. Cho, W., Choi, S., Park, D. K., Shin, I., and Choo, J. (2019). Image-to-image translation via group-wise deep whitening-and-coloring transformation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp.10639-10647).

[20]. Animesh K. and Oliver W. (2020). MSG-GAN: Multi-scale gradients for generative adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 7799–7808.

[21]. Guarnera L., Giudice O., and Battiato S, (2020). Deepfake detection by analyzing convolutional traces. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 666-667.

[22]. Moon T.K. (1996). The expectation –maximization algorithm. IEEE Signal Processing Magazine, 13(6), pp.47-60.

[23]. Liu, Z., Luo, P., Wang, X., and Tang, X. (2018). Large-scale celebfaces attributes (celeba) dataset. In Proceedings of the IEEE International Conference on Computer Vision (pp. 3730-3738).

[24]. Hsu, C. C., Zhuang, Y. X., and Lee, C. Y. (2020). Deep fake image detection based on pairwise learning. Applied Sciences,10(1), 370.

[25]. Chopra, S. (2005). Learning a similarity metric discriminatively, with application to face verification. In IEEE Conference on Computer Vision and Pattern Recognition, pp. 539-546.

[26]. Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger,K. Q. (2017). Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision andPattern Recognition (pp. 4700-4708).

[27]. Li L., Bao J., Zhang T., Yang H., Chen D., Wen F., and Guo B., (2020), Face X-ray for more general face forgery detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5001-5010).

[28]. Xuan, X., Peng, B., Dong, J., and Wang, W. (2019). On the generalization of GAN image forensics. arXiv preprint arXiv:1902.11153.

[29]. Zhao T., Xu X., Xu M., Ding H., Xiong Y. and Xia W., (2021),Learning self-consistency for deepfake detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision . pp. 15023-15033.

[30]. Rossler A., Cozzolino D., Verdoliva L., Riess C., Thies J., Niebner M., (2019). Faceforensics++:Learning to detect manipulated facial images. In proceedings of the IEEE/CVF International Conference on Computer Vision. Pp:1-11.

[31]. Dolhansky, B.; Bitton, J.; Pflaum, B.; Lu, J.; Howes, R.; Wang, M.; Canton Ferrer, C. (2020). The deepfake detection challenge dataset. arXiv 2020, arXiv:2006.07397.

[32]. Sitong L., Zhichao L., Siqi G. , Liang X., (2022). Block shuffling learning for deepfake detection.Computer Vision and Pattern Recognition. arXiv:2202.02819v1.

[33]. Li, Y.; Yang, X.; Sun, P.; Qi, H.; Lyu, S.C.D. (2020). A large-scale challenging dataset for DeepFake forensics. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19; pp. 14–19.

[34]. Afchar, D., Nozick, V., Yamagishi, J., and Echizen, I. (2018). MesoNet: a compact facial video forgery detection network. In 2018 IEEE International Workshop on Information Forensics and Security (WIFS) pp. 1-7 . IEEE.

[35]. Viola P. and Jones M. (2021). Rapid object detection using a boosted cascade of simple features. In Computer Vision and Pattern Recognition. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on, volume 1, pages I–I.IEEE.

[36]. Aya I., Marwa E., Mervat S. Z., and Kamal E., (2021). A new deep learning-based methodology for video deepfake detection using Xgboost. Sensors,21, 5413.

[37]. Chen T., Guestrin C. (2016), Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Pp. 785-794.

[38]. Joseph R., Santosh D., Ross G., and Ali F.(2016). You Only Look Once: Unified, Real-Time Object Detection. In proceedings of the IEEE Conference in Computer Vision and Pattern Recognition, pp 779-788.

[39]. Amerini I., Galteri L., Caldelli R., Del Bimbo A.(2019) Deepfake Video Detection through Optical Flow Based CNN Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).

[40]. Tran V.N, Lee S.H, Le H.S, Kwon K.R. (2021) High Performance DeepFake Video Detection on CNN-Based with Attention Target-Specific Regions and Manual Distillation Extraction.. Applied Sciences. 11(16):7678. https://doi.org/10.3390/app11167678.

[41]. Goyal K., Agarwal K., Kumar R. (2017). Face detection and tracking: Using OpenCV. In Proceedings of the 2017 International conference of Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, Volume 1, pp. 474–478.

[42]. Zhang K., Zhang Z., Li Z., Qiao Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process. Lett. 23, 1499–1503.

[43]. Baltrusaitis T., Zadeh A., Lim Y.C., Morency L.P.(2018). Openface 2.0: Facial behavior analysis toolkit. In Proceedings of the 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), Xi'an, China, 15–19 May 2018; pp. 59–66.

[44]. Szegedy C., Vanhoucke V., Ioffe S., Shlens J., Wojna Z. (2016). Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–2 July 2016; pp. 2818–2826.

[45]. Howard A.G., Zhu M., Chen B., Kalenichenko D., Wang W., Weyand T., Adam H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv:1704.04861.

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

[47]. Deressa W.,and Solomon A., (2021). Deepfake Video Detection Using Convolutional Vision Transformer. Computer Vision and Pattern Recognition, arXiv:2102.11126v3.

[48]. Guera D., and Delp E. J. (2018, November). Deepfake video detection using recurrent neural networks. In 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS) (pp. 1-6). IEEE.

[49]. Hochreiter S. and Schmidhuber J. (1997). Long Short-term memory. Neural Computation, 9(8), pp: 1735-1780.

[50]. Laptev I., Marszalek M., Schmid C., and Rozenfeld B. (2008). Learning realistic human actions from movies. Proceedings of the IEEE Conference on computer Vision and Pattern Recognition, pages 1–8, June 2008. Anchorage, AK.

[51]. Abdul Jamsheed V., and Janet B. (2021). Deep fake video detection using recurrent neural networks. International Journal of Scientific Research in Computer Science and Engineering. Vol.9(2). Pp:22-26.

[52]. Priti Y., Ishani J., Jaiprakash M., Vibhash C. and Gargi Kh., (2021). International Conference on Emerging Technologies: AI, IoT, and CPS for Science Technology Applications, 06-07.

[53]. Saining X., Ross G., Piotr D., Zhuowen T., Kaiming H. (2017). Aggregated Residual Transformations for Deep Neural Networks. https://arxiv.org/abs/1611.05431v2.

[54]. Li, Y., Chang, M. C., and Lyu, S. (2018, December). Inictu oculi: Exposing AI created fake videos by detecting eyeblinking. In 2018 IEEE International Workshop on InformationForensics and Security (WIFS) (pp. 1-7). IEEE.

[55]. Donahue J., Anne Hendricks L., Guadarrama S., Rohrbach M., Venugopalan S., Saenko K., and Darrell T. (2015). Long-term recurrent convolutional networks for visual recognition and description. in CVPR, 2015, pp.2625–2634.

[56]. Simonyan K. and Zisserman A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[57]. Song F., Tan X., Liu X., and Chen S. (2014). Eyes closeness detection from still images with multi-scale histograms of principal oriented gradients. Pattern Recognition, vol. 47, no. 9, pp. 2825–2838.

[58]. .Shahroz T., Sangyup L., and Simon S. W. (2021). One Detector to rule them all:Towards a General Deepfake Attack Detection Framework. Computer Vision and Pattern Recognition. https://arxiv.org/abs/2105.00187v1.

[59]. Daichi Z., Chenyo L., Fanzhao L., Dan Z., and Shiming G. (2021). Proceeding of the Thirtieth International Joint Conference on Artifitial Intelligence (IJCAI).

[60]. Yipin Z. and Ser-Nam L., (2021). Joint Audio-Visual Deepfake Detection.IEEE/CVF International Conference on Computer Vision (ICCV), pp. 14780-14789, doi: 10.1109/ICCV48922.2021.01453

[61]. Shruti A., Hany F., Ohad F.and Maneesh A., (2020). Detecting Deep-Fake Videos from Phoneme-Viseme Mismatches. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). DOI 10.1109/CVPRW50498.2020.00338.

[62]. Wang Y., and Dantcheva A. (2020). A video is worth more than 1000 lies. Comparing 3DCNN approaches for detecting deepfakes. FG 2020 - 15th IEEE International Conference on Automatic Face and Gesture Recognition, Nov 2020, Buenos Aires / Virtual, Argentina. hal-02862476.

[63]. Andreas R., Davide C., Luisa V., Christian R., Justus T., Matthias N. Faceforensics: A large-scale video dataset for forgery detection in human faces, CoRR abs/1803.09179.

[64]. Yuezun L., Ming-Ching C., Siwei L. (2018) In ictu oculi: Exposing AI created fake videos by detecting eye blinking, in: IEEE International Workshop on Information Forensics and Security (WIFS), IEEE, pp. 1-7.

[65]. Korshunov P., Sebastien M. Deepfakes: a new threat to face recognition? assessment and detection, arXiv preprint arXiv:1812.08685.

[66]. .Liming, L. Ren, W. Wayne, Q. Chen, L. Chen Change. (2020) Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2889-2898.