International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

A Survey an Optimized Dense CNN Model for Recognizing Deepfake Images

Mallikarjun Gachchannavar[1], Dr. Naveenkumar J.R.[2], Radha Velangi[3]

[1]Lecturer in Electronics and Communication Engineering Government Polytechnic, Belagavi
[2]Professor and HOD in Electronics and Communication Engineering, Srinivas Institute of Technology, Mukka, Srinivas University, Mangaluru
[3]Lecturer in Electronics and Communication Engineering Government Polytechnic, Belagavi

Abstract

Deepfake detection relies on a deep learning model. Deepfake content is generated using artificial intelligence and machine learning techniques to swap one person's face with another's in images. These altered images are certain to have a significant impact on society. Deepfakes utilize advanced technologies such as machine learning (ML) and deep learning (DL) to develop automated techniques for generating deceptive content. The model is trained using the DFDC (Deepfake Detection Challenge) dataset, which includes 1, 00,000 videos comprising both real and fake content. This survey examines recent progress in deep learning techniques for detecting deepfake images, highlighting their growing significance in today's digital landscape. The paper outlines the issues arising from deepfakes, explores the techniques for generating deceptive content, and analyzes various deep learning models employed in deepfake detection, such as convolutional neural networks (CNN's), MTCNN (Multi-task CNN), and Facial Landmark. It emphasizes the ongoing battle between the progression of deepfakes and the evolution of detection methods, underscoring the need for advanced and flexible neural network architectures to effectively curb the dissemination of deceptive information.

Keywords: CNN **(**Convolutional Neural Networks), DFDC, MTCNN, and Facial Landmark.

1. Introduction

[1] Deepfakes, mainly created using deep learning models such as auto encoders and generative adversarial networks (GANs), have raised serious concerns because of their potential for misuse. By harnessing large image and video datasets, these AI-based methods create convincing facial expressions and movements, often aimed at public figures like celebrities and politicians. The danger of deepfakes goes beyond producing fake adult content; it includes altering political events, swaying elections, disseminating misinformation, and even deceiving military intelligence. If misused, this technology can produce misleading content, including fabricated speeches by global leaders, fake satellite imagery, and even altered historical records. Despite these concerns, deepfakes also offer beneficial applications, such as improving visual effects, creating digital avatars, generating voices for those without speech abilities, and enhancing entertainment experiences.

[2] While deepfakes have promising benefits, their widespread use necessitates strong measures to prevent the use necessitates strong measures to prevent the risks of misuse. However, the troubling

reality is that the harmful uses of deepfakes greatly over shadow their positive applications. The development of advanced deep neural networks, coupled with the widespread availability of data, has made altered images and videos almost indistinguishable from genuine ones, challenging both human perception and sophisticated computer algorithms. The simplicity of manipulation has greatly reduced the difficulty of producing convincingly realistic content, often needing just a single photo or short video clip of a person. Notably, recent developments have made it possible to create convincing deep fakes using just a single static image. As a result, the threat of deep fakes now affects not only public figures but also ordinary individuals.

1.1 Deep Learning for Deepfake Detection:

DL is a particular area of machine learning that centers on learning data representations via artificial neural networks with multiple layers. The basic building blocks of a neural network is the neuron, which takes in inputs, processes them through weighted connections, and uses an activation function to generate an output. Deep learning comprises a range of architectures, such as feedforward neural networks, recurrent neural networks (RNNs), convolutional neural networks (CNNs), and more. Deep learning techniques are often used to detect and identify deepfake content because of their capacity to analyze and recognize patterns in complex data like images and videos. **[3]** Some commonly used deep learning methods for detecting deepfakes include:

1.1.1 CNN

A convolutional neural network (CNN) is a type of neural network specifically designed to learn feature engineering by optimizing its filters. This regularization technique enables CNNs to automatically extract relevant features from input data, eliminating the need for manual feature engineering. As illustrated in fig1, CNNs are composed of convolution are widely used in tasks like fake photo detection and object recognition because they are highly effective at extracting features by employing linear algebra principles, especially matrix multiplication to recognize patterns in images.

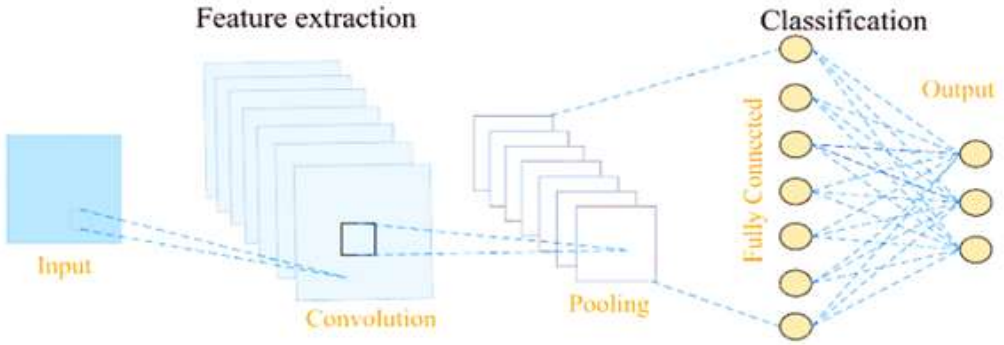

Figure 1. CNN Architecture

The basic framework of a CNN model comprises three types of layers: convolutional, pooling, and fully connected. Figure 1 show the basic structure of a CNN model, where the convolution layer is responsible for feature extraction. During the convolutional operation, a set of numbers (kernel) is applied to the input (tensor) to generate the feature map. The process of creating a feature map involves an element-wise multiplication between the kernel and the input tensor, with the resulting values summed to produce each element of the feature map. The kernel moves across all elements of the input

tensor to generate the elements of the features map for that kernel. Using different kernels in the convolution operation can generate multiple feature maps.

1.1.2 Recurrent Neural Networks (RNNs):

RNNs are employed to handle sequential data and can analyze temporal patterns across video frames, assisting in the detection of irregularities or inconsistencies in the frame sequence that could indicate manipulation.

1.1.3 Generative Adversarial Networks (GANs):

While GANs are commonly used to create deepfakes, they are also utilized for detecting them. GAN-based detection models are trained to distinguish between genuine and manipulated content by spotting discrepancies in generated images or videos.

1.1.4 Capsule Networks:

These networks are built to manage hierarchical relationships among features, which can enhance the understanding of spatial hierarchies and assist in detecting unnatural distortions in deepfake images or videos.

1.1.5 Siamese Networks:

Siamese networks, employed for one-shot learning, can detect discrepancies between original and altered content by learning feature representations that reveal differences.

1.1.6 Deep Feature Extraction:

Pre-trained models like VGG, ResNet, or EfficientNet extract deep features from images or videos, which are then used by traditional machine learning classifiers to identify anomalies in the data.

1.2 Dataset

The dataset used for training the model is DFDC (Deepfake Detection Challenge Dataset) **[4]** which is publicly accessible on kaggle. The DFDC dataset consists of 100000 videos, each approximately 10 seconds long, with images extracted from a randomly chosen subset of these videos. Because there were more fake videos than real ones, the dataset was balanced by randomly removing the excess fake images after frames were extracted at 1 fps.

1.3 MTCNN (Multi-task CNN)

The MTCNN network, known for its high detection accuracy, lightweight design, and real-time performance, is used for face recognition our face recognition process is therefore divided into two main steps: face detection and face recognition. Initially, MTCNN is used for face detection to determine accurate face coordinates. The processing flow of MTCNN begins with repeatedly resizing the test image to generate an image pyramid. Next, the image pyramid is input into P-Net to generate numerous candidate faces. The candidates identified by P-Net are then refined by R-Net. Once R-Net eliminates many candidates, the images are passed to O-Net, which then outputs the accurate box coordinates. Unlike deepfake, FaceNet keeps face alignment, skips the feature extraction steps, and uses CNNs for end-to-end training directly after the face alignment. The “Real and Fake Face-Detection” dataset was used to train the three models, with a learning rate of 0.001 and training period of 10 epochs. As a result, the accuracy on the test set was determined by evaluating the testing dataset. To increase the dataset size, all original images were flipped both vertically and horizontally, tripling the amount of data.

2. Literature Review

In the paper by **Ali s.et al (2022),** a new image forgery detection system was proposed that utilized neural networks, with a particular emphasis on the Convolutional Neural Network (CNN) architecture. The method utilized variations in image compression by applying the differences between original and recompressed images to train the model. This approach successfully identified prevalent types of image forgeries, including splicing and copy-move manipulations. The authors identify a major gap in existing CNN-based forgery detection methods, which generally specialize in identifying a single type of forgery.

The proposed system aims to address this limitation by efficiently detecting a wider range of novel forgeries within images. The primary innovation is in leveraging the differences between an image's original and recompressed versions as signals for detecting forgeries. This indicates that the model is intended to identify artifacts and irregularities introduced during the recompression process, which can be indicative of manipulations. The authors stress the lightweight design of the proposed model, emphasizing its focus on computational efficiency and real-time processing. This is important for practical applications, especially in cases where rapid and efficient forgery detection is necessary. The experiments produced very promising result, revealing an impressive overall validation accuracy of 92.23% within the iteration limits. **[5]**

The study by **Suganthi et.al. 2022** employed the FF-LBPH DBN (Fisherface Linear Binary Pattern Histogram with DBN Classifier) technique for deepfake image detection, highlighting its impressive speed and its effectiveness in distinguishing between real and fake images. The proposed approach adopts a hybrid methodology, starting with the use of Fisherface with Local Binary Pattern Histogram (FF – LBPH) for face recognition, focusing on reducing dimensionality in the face space. Following this, a Deep Belief Network (DBN) using Restricted Boltzmann Machines (RBM) is utilized for detecting deep fakes. This approach aim to boost the accuracy and efficiency of the detection process, offering potential to prevent underserved defamation from manipulated images. The authors point out that inaccuracies and lengthy processing times are major issues in existing deepfake detection techniques. The FF-LBPH model attained remarkable accuracy rates, achieving 98.82% on the CASIA-WebFace dataset and 97.82% on the DFFD dataset. The research concluded by emphasizing the exceptional performance of FF-LBPH in detecting and analyzing deepfake face images. **[6]**

Yogesh Patel et.al carried out a study on deepfake image detection employing a dense CNN architecture. The dataset utilized in the study is sourced from the deepfake images detection and reconstruction challenge, including real images from CelebA and FFHQ, as well as deepfake images from GDWCT, AttGAN, STARGAN, StyleGAN, and StyleGAN2. They introduced a D-CNN model for binary classification to identify deepfake images. The augmented CNN model is integrated with the D-CNN model to extract deep features from input images using convolutional layers. Once convolution operations are performed on the images, they can be used to classify the input images as either fake or real. The proposed model attained 97.2% accuracy on the test dataset. The models performance declines when applied to other existing models, such as MesoNet and MesoInception networks, on the CelebDF dataset. **[7]**

In their recent paper, **El-Gayar MM et.al.** Introduce a new method for deepfake detection using Graph Neural Network (GNNs). GNNs are adept at modeling relationships between data points. When applied to deepfakes, this capability translates into capturing the complex interactions between facial features in videos. By examining these relationships, GNNs can detect inconsistencies that suggest manipulation,

such as unnatural connections between facial landmarks (e.g., eyes and nose). The paper also investigates combining GNNs with other models, like CNNs, through fusion strategies, which could further improve deepfake detection performance. **[8]**

Hanqing Zhao et.al has redefined deepfake detection as a fine grained classification problem, offering a novel approach to the field. They further introduced an innovative multi-attention network architecture crafted to capture local discriminative feature from various face-focused regions. The datasets employed are FaceForensics ++, CelebDF, and DFDC. The proposed architecture splits the single attention-based structural networks into multiple regions, improving efficiency in capturing local features. They utilized local attention pooling to capture textual patterns. The implementation of the multi-attention framework leads to significant improvement across various datasets using comprehensive metrics. **[10]**

Tran VN et.al presents Meta deepfake detection (MDD) **[11]** in this paper as a solution to this challenge, with the goal of improving performance, especially in unexplored domains. MDD leverages meta-learning, a strategy that allows for the acquisition of transferable knowledge from different source domains. Through meta-learning, the model becomes skilled at adjusting to unfamiliar, unseen domains without needing frequent updates. Essential elements of MDD involve meta-splitting, where source domains are divided into meta-train and meta-test sets to reflect real-world domain shifts.

Yuval Nirkrin et.al has introduced a method for detecting deepfake images by identifying mismatches between faces and their context. The datasets employed for this method are FaceForensics++,CelebDF, and DFDC. Their proposed method incorporates two networks: one for identifying the face and its surrounding area, and another for detecting facial landmarks using an xception network that considers the context of the face. This new method greatly exceeds the performance of the baseline xception model. Nonetheless, it might not perform as well on images with low contrast and blurred features. **[12]**

Employing a multi-task learning strategy, **Nguyen et.al. (2019)** carefully developed a CNN in their study to concurrently detect altered images and videos as well as to precisely locate the edited regions within them. The method adopted by the network allowed information acquired from one task to support the other, fostering a cooperative relationship that enhanced the performance of both detection and localization tasks. The application of a semi-supervised learning approach significantly improved the networks ability to generalize across a variety of datasets. The core approach involved leveraging the difference between images original and recompressed versions to train the deep learning model. This strategic selection of training data was closely aligned with the specific type of forgery being examined, offering a focused investigative approach. **[13]**

The goal of the framework proposed by **Ambica Ghai et. al. (2021)** was to detect forged images altered using copy-move and splicing techniques. Utilizing an image transformation technique helped isolate relevant features essential for efficient network training. Following this, a pre-trained customized CNN was employed for training on publicly available benchmark datasets. The selection of a CNN reflects a dependence on a deep learning architecture particularly suited for image-related tasks. Performance assessment was performed on the test dataset, utilizing different metrics to evaluate the models effectiveness. The research confronts the widespread issue of online misinformation underscoring the effect of manipulated content on decision-making. Emphasizing the important role of images in complementing textual information and the growing prevalence of image manipulation, the study aims to create a deep learning-based framework for detecting image forgeries. **[14]**

Sohail Ahmed Khan et.al have developed a hybrid transformer network for detecting deepfake images. The dataset utilized are FaceForensics++ and DFDC. Two CNN architectures, XceptionNet and

effcientNet-B4, are utilized for feature extraction and BERT- style transformer is then applied to integrate the features. The model may struggle with unseen data that features different types of forgery techniques applied to images. **[15]**

A multi-attention network is a neural network architecture that combines multiple attention mechanisms to capture varied and detailed information from input data. In deep learning, attention mechanisms enable networks to concentrate on particular parts or aspects of the input, allowing for more efficient processing of complex data. A network proposed by **Zhao et.al 2020 [16]** utilizes a deep learning architecture that integrates multiple spatial attention heads, enabling the model to focus on various local regions of the images. The main emphasis is on improving the detection of deepfake content by taking into account the suitable and localized differences between real and fake images. The paper challenges the common practice of handling deepfake detection as a simple binary classification task and introduces a new multi-attentional deepfake detection network.

The framework proposed by **Ambica Ghai et. al. 2021 [17]** aimed to detect forged images that had been manipulated using copy-move and splicing techniques. An image transformation technique was used to isolate relevant features essential for effective network training. Following this, a pre-trained custom CNN was employed for training on publicly available benchmark datasets. Opting for a CNN demonstrates a reliance on deep learning architecture that is particularly effective for image-based tasks. The model's performance was assessed on the test dataset using various parameters. The research tackles the widespread problem of online misinformation highlighting the impact of manipulated content on decision-making. Highlighting the importance of images in augmenting textual information and the increase in image manipulation, the study aims to develop a deep-learning-based framework for detecting forged images.

Li et al. [18] introduced a new method for detecting fake or altered faces in unmodified images or videos. They focused on a key aspect of human facial behavior eye blinking rate-to authenticate physiological single that are often missing or inaccurately represented in synthetic fake videos, as shown in figure1.

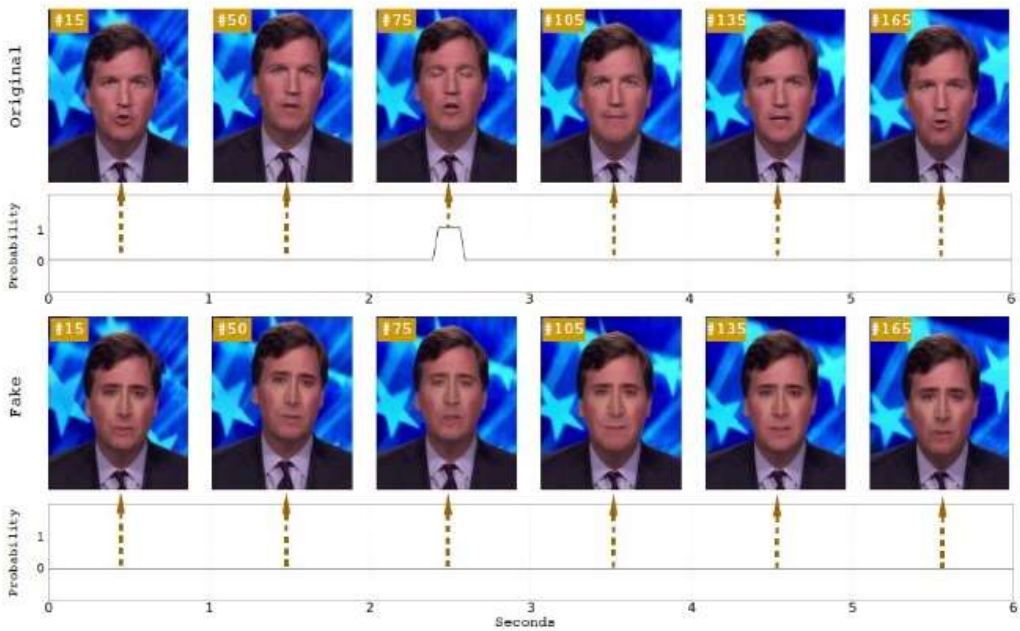

Figure 2. Eye-blinking detection on an original video (**Top**) and a DeepFake-generated video (**Bottom**).

This study analyzed the eye blinking rates in original videos and compared them with those in DeepFake videos. The final results showed that the proposed DeepFake detection method could identify a synthetic or fake video by detecting abnormal or anomalous eye blinking rates. Figure 1 illustrates a frame-by-frame analysis of eye blinking in both an original video and a DeepFake video. The authors measured the average duration between consecutive eye blinks and the average time required to notice eye blinks to detect wheter the videos were real or fake. This method involves two stages: (a) detecting faces in images or frames, identifying facial landmarks, aligning faces, and extracting the eye region and (b) The features from the first stage are input into a long-term recurrent convolutional network (LRCN) to track the number of eye blinks.

Afchar et al. [19] concentrated on examining the mesoscopic properties of images by employing detection systems based on a deep learning approach. They incorporated two distinct activation functions and introduced two detection methods, Meso-4 and MesoInception-4 to distinguish between fake and real videos or images.

In **Meso-4,** they implemented four sequential layers of convolution and pooling, followed by a dense network with one hidden layer that utilized the Rectified Linear Unit (ReLU) activation function to enhance generalization.

In the **MesoInception-4** architecture, the authors substituted the initial two convolution layers with inception models and subsequently evaluated the model on the DeepFake and Face2Face datasets. The results indicated a very high detection success rate, with 98% accuracy for the deepfake dataset and 95% for the Face2Face dataset.

Hinton et al. [20] identified key limitations of convolutional neural networks (CNNs) and proposed a new capsule architecture as a foundational solution. **Nguyen et al. [21]** built on the concept of capsule architecture and expanded their research to detect various types of images and video forgeries, including replay attacks. The proposed system integrated advanced deep convolutional neural networks (DCNNs) and evaluated its performance against other benchmarking methods. This method is frequently utilized with dynamic routing algorithms and expectation maximization routing algorithms.

Rossler et al. [22] introduced an automated pipeline for detecting fake faces in images or videos. This approach involved using a tracking algorithm to identify and track human faces, which were then fed into different classifiers to determine if forgery was present in the videos. The proposed method utilized four deepfake datasets-deepfakes, Face2Face, Faceswap, and NeuralTextures- along with a pristine dataset to evaluate precision.

Yuval Nirkrin et.al proposed a method for detecting deepfake images by identifying inconsistencies between faces and their surrounding context. The data sets used for this method include FaceForensics++, CelebDF, and DFDC datasets. There method utilizes two networks: one for detecting the face and its surrounding region, and another for identifying facial landmarks based on an xception network that considers the context of the face. This new method surpasses the baseline xception model by a considerable margin. However, it may struggle with images that have low contrast and blurred features.**[23]**

Sohail Ahmed Khan et.al [24] introduced a hybrid transformer network for detecting deepfake images, using the FaceForensics++ and DFDC datasets. The method utilizes two CNN architectures, XceptionNet and EfficientNet-B4, for feature extraction, and then applies a BERT-Style transformer to integrate the features. The model may struggle with unseen data that exhibits various styles of forgery techniques applied to the images.

Ali Raza et.al introduced a new deep learning technique for detecting deepfake images, using a deepfake dataset publicly available on kaggle. The novel DFP approach combines VGG16 with convolutional neural architecture. It uses hybrid layers from both to build the overall architecture. The DFP approach suprassed other state-of-the-art methods in performance. However, it struggles to generalize across different types of images generated by various techniques. **[25]**

Asad Malik et.al reviewed deepfake detection methods for human face images and videos, examing various algorithms that utilized different datasets. The datasets used include Celeb-DF, deep forensics, wide deepfake dataset, and open forensics dataset. A verity of algorithms and CNN models were employed for a thorough analysis. **[26]**

Korshunov et al. [27] generated DeepFake videos using the VID-DIMIT dataset. The employed open-source GAN-based software to create these deepfakes and highlighted how training and blending parameters effect the quality of the videos, both low and high, with different sets of tuned parameters. For each of the 320 subjects, they produced two video versions: one using a low quality (64X64) GAN model and the other using a high quality (128X128) model. They also showed that state-of-the-art VGG and FaceNet- based Face Recognition algorithms are susceptible to deepfake videos, filing to differentiate these videos from the originals, with a false acceptance rate of up to 95.00%.

Korshunov et al. [28] also assessed baseline face-swap detection algorithms and discovered that the lip-sync-based method was inadequate for identifying mismatches between lip movements and speech. They also validated that image quality measures combined with a support vector machine (SVM) classifier can identify high-quality deepfake videos, with an equal error rate of 8.97%.

Agarwal and Varshney [29] created a statistical model using hypothesis testing to detect face-swapped content or fraudulent alterations in images. In this study, the authors took into account a mathematical bound value related to the error probability when detecting genuine versus GAN-generated images.

Lyu [30] pointed out the major difficulties in detecting DeepFakes in high-quality or high-definition synthesized videos and audio recordings. The author expressed serious concern about a significant drawback of current deepfake generation methods, which struggle to accurately map color shades for hair in reaction to the human face.

Kumar et al. [31] utilized multiple deep learning approaches and compared their results in the context of deepfake classification through metrics learning. The authors utilized a multitask cascaded convolutional neural network (MTCNN) to extract faces from images or videos. The MTCNN includes three networks: (a) a proposal network, (b) a refine network, and (c) output networks, which apply non-max suppression to remove overlapping boxes and produce bounded faces.

The xception architecture was utilized for transfer learning; with sequence classification performed using LSTM, in combination with 3D convolution and a triple network. A triplet network, combined with metric learning, was used to develop an approach that counts the number of frames in a given video clip. The realism factor had to be assessed if the number of frames was less than the actual count when compared to the original video. This study examined three types of triplet-generation methods: easy triplets, semi-hard triplets, and hard triplets. These methods are based on the distances between the anchor, positive, and negative embedding vectors.

The proposed detection Architecture, illustrated in Figure 3, employs XceptionNet for the entire process along with the MTCNN. In the initial phase, the facial landmark model is used to detect and extract facial features, generating a feature space of 512-dimensional embedding vectors for each face. The generated feature space is then input semi-hard triples, which differentiate between fake and pristine

frames using triplet loss. During validation, this method attained an AOC score of 99.2% on Celeb-DF and an accuracy of 99.71% on highly compressed neural textures **[32].**

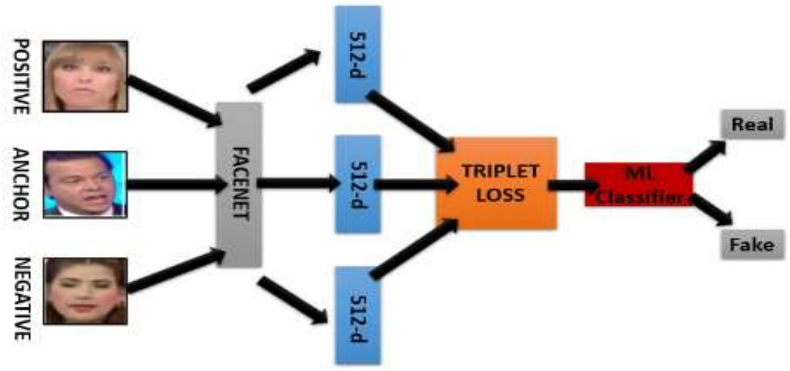

Figure 3 Triplet architecture used for clustering and classification of real videos embedding vectors.

Mittal et al. [33] presented a method that combines a convolutional neural network with a recurrent neural network, enabling the extraction of crucial temporal features from faces to detect manipulated or synthesized images. A Gated Recurrent Unit (GRU), combined with a weighting mechanism and automatic face weighting (AFW), was employed to automatically select the most reliable frames for detecting forged faces.

Figure 4 illustrates the complete execution flow of the proposed detection Architure for determining the authenticity of genuine or fabricated videos. The process, facial features are detected and extracted from multiple frames using MTCNN. Following the detection of face regions, a binary classifier is trained using efficientNet-b5 to extract features that differentiate between real and fake faces. Finally, the prediction for classifying realism or fakeness is determined by combining AFW with GRU. The authors trained and assessed the proposed method using the deepfake detection challenge (DFDC) dataset, achieving a log-likelihood error of 0.321.

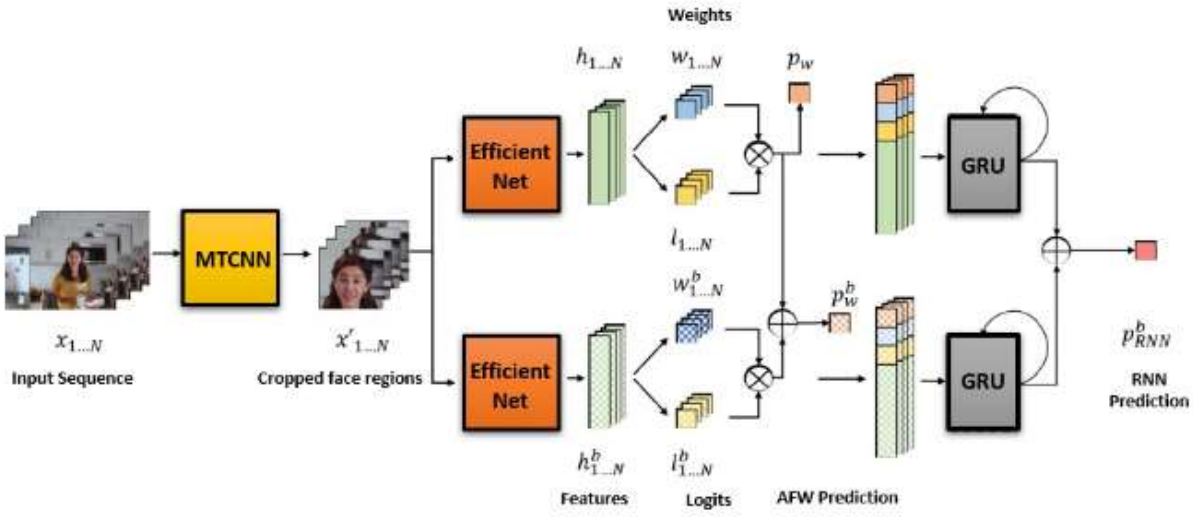

Figure 4. Extraction of the face from the frame using MTCNN algorithm.

Kawa and Syga [34] proposed two Deep Fake detection models that delivered higher accuracy with reduce computational expense. In their first approach, they improved the MesNet model by integrating a new activation function called the pish activation function. MesNet employed a convolutional neural network, which came in two forms: Mes04 and Mesolnception-4.

Chugh et al. [35] introduced a technique that uses the modality dissonance score (MDS) to identify forgery in Deep Fake videos by analyzing the discrepancies between audio and visual elements. They employed contrastive loss to assess how similar features are between audio and video, and used entropy loss to differentiate between the audio and video modalities.

Kaur et al. [36] applied a deep depth-based convolutional long short- term memory (C-LSTM) model to sequential video frames for Deep Fake detection. They retrieved frames from the Deep Fake video using "OPENCL" which combines features from the source frames with those in the target frames of the video. The model consists of two-tier deep temporal C-LSTM, where the first tier extracts frames from the manipulated video and passes them to the C-LSTM model for detecting Deep Fakes.

Rahul et al. [37] created a method that leverages common features of altered video clips to analyze facial recognition. This approach uses a sandwich technique, where manipulated videos are converted into frames and then analyzed by the MTCNN to extract facial features with the Mobile Net model. The pre-trained Mobile Net serves as the input, and transfer learning is applied to this model to classify videos as either fake or genuine. Tested on the Face Forensic dataset, this method achieved an average accuracy of 85% in detection.

Wubet [38] applied a CNN that combines ResNet and VGG-16 for eye state classification, along with long short-term memory (LSTM) for sequence analysis. The study evaluated whether videos from the UADFV dataset were genuine or counterfeit by counting eye blinks over time, using the eye aspect ratio to gauge the dimensions of open and closed eyes, as shown in Figure5.

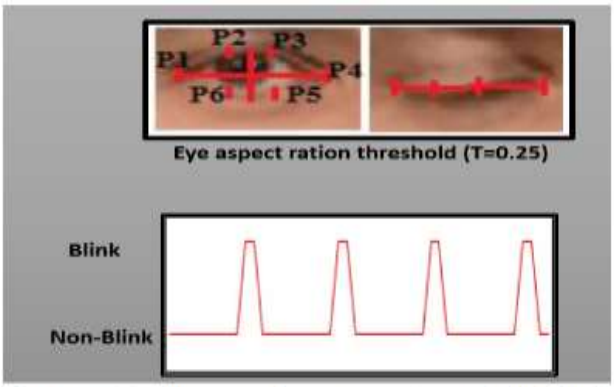

Figure 5. Coordinates to detect eye regions and the blinking of the eye.

In the figure, point's p2, p3, p5 and p6 are used to measure the vertical height of the eye, while point's p1 and p4 measure the horizontal width. These measurements are essential for identifying whether the eyes are open or closed. The research used the average rate of human blinking as a benchmark to detect and count blinks and intervals, considering that humans typically blink every 2 to 10 seconds, with each blink tasting between 0.1 and 0.4 seconds. Based on these measurements, the study was able to classify video as fake or real. It identified 184 blinks per minute in real videos and 428 blinks per minute in fake videos, an accuracy of 93.23% for real videos and 98.1% for fake ones.

Pishori et al. [39] optimized eye blink detection strategies for Deep fake detection by developing a three-phase model that integrates convolutional LSTM, eye blink measurement, and grayscale histograms. This model utilized a bland of CNN and RNN to monitor eye blinks and relied on the OpenCv library to located facial landmarks in images or video frames.

Hussain et al. [40] proposed an innovative SOA technique for evading Deep Fake detectors if the attacker has full or partial insight into the detection system. They developed adversarial examples for each frame of a fake video and combined them to create an adversarial altered video that was

misclassified as real by the Deep Fake detectors, XceptionNet and Meso Net. They explores both white-box and black-box attack strategies for video classification. As Deep Fakes become more prevalent, various researches and esteemed institutions have carried out extensive reviews, highlighting different models or systems for detecting fake images or videos. Notable contributions include vakhshiteh et al.[41], Nguyen et al. [42], Mirsky and lee [43], Tolosana et al. [44], Sohrawardi et al. [45], and Verdoline [46], who have examined the advantages and drawback of Deep fake technologies and provided in-depth analysis of detection mechanisms.

Guera and Delp [47] introduced a two-step analysis method, beginning with a convolutional Neural Network (CNN) to extract features from individual frames. These features were then input into a recurrent Neural Network (RNN) to assess whether the videos are genuine or fake. They successfully identified temporal inconsistencies between frames resulting from face swapping. In the CNN stage, they used Inception-v3 with an additional fully connected layer ay the networks top. During the LSTM processing stage, a softmax layer was employed to evaluate inconsistencies within and between frames, which are typically introduced by face swapping or deep fake manipulations. The LSTM was followed by a fully connected layer with 512 units and a dropout rate of 0.5%.

Guarnea et al. [48] introduced a method for detecting Deep Fakes of human face by uncovering forensic traces hidden in images using Expectation Maximization (EM) algorithms. The EM algorithms were employed to extract a set of local features from the images, and the approach was validated through tests using a naïve classifier on five different architectures (GDWCTS, STARGAN, ATTGAN, STYLEGAN, and STYLEGAN2) against the CELEBE datasets.

Huang et al. [49] introduced a fake polisher technique that utilizes a post-processing shallow reconstruction approach, designed to bypass existing state-of-the-art (SOA) detection methods without requiring prior knowledge of the GAN. Current GAN-based images generation methods often leave artifact patterns due to inherent limitations, which the authors aimed to address. They proposed method that effectively identify these artifact patterns and minimize them in synthesized images. A multi-attention network is a neural network framework that integrates multiple attention mechanisms to extract varied and detailed information from input data. In deep learning, attention mechanisms help networks focus on specific element of the input, improving the handling of complex data.

Zhao et al. [50] proposed such a network, which utilizes a deep learning architecture with multiple spatial attention heads, enabling the model to concentrate on different local regions of images. The primary objective is to improve deep fake detection by focusing on subtle and localized distinction between real and fake images. The paper critiques the common approach of treating deep fake detection as a simple binary classification task and suggests a novel multi-attentional deep fake detection network.

Gandhi, A et.al 2020 [51] introduced a method where adversarial perturbations were used to trick deep fake detectors, causing detection accuracy to drop to below 27% on altered deep fakes. To enhance detection resilience, two strategies were investigated: Lipchitz regularization, which improved accuracy by 10% in black-box scenarios, and Deep image prior (DIP), which achieved 95% accuracy in detecting altered deep fakes while maintaining 98% accuracy in other instances on a set of 100 images. These approaches were designed to bolster deep fake detection against adversarial attacks, showcasing significant advancements in defensive techniques. Additionally, the research utilized VGG and ResNet architectures, well-known convolutional neural network models, to improve the identification of unique features in images, further advancing the detection of deep fakes in the face of adversarial challenges.

Another study by **Nguyen, H et.al. [52]** Conducted an in-depth investigation into several key areas, such as Replay attack detection, face swapping detection, reenactment detection, and computer-generated image detection. This through analysis highlighted the flexibility of capsule networks beyond traditional computer vision applications. Moreover, the research emphasized the benefits of incorporating random noise during the training process, proving its effectiveness in various contexts.

Table 1. Survey table

Study	Year	Network Type	Success Metrics	Key Aspects
Ali.S.et.al(2022) **[53]**	2022	CNN	Effective in detecting image forgeries	Lightweight model, leverages differences between original and Recompressed images for training
Suganthiet.al. (2022) **[54]**	2022	FF-LBPH with DBN	Impressive accuracy rates	Hybrid methodology, utilizes FF-LBPH for face recognition and DBN for deepfake detection
Lakshmanan Nataraj et.al. (2019) **[55]**	2019	CNN with Co-occurrence Matrices	Effective integration of matrices	Hybrid approach, combines co- occurrence matrices with CNN
Luca Guarnera et.al. (2020) **[56]**	2020	EM Algorithm	Distinguishing deep fake architectures	Leveraging EM algorithm for feature extraction in deepfake creation
Ambica Ghai et.al. (2021) **[57]**	2021	CNN	Addressing image forgery detection	Image transformation technique For relevant feature isolation
Zhao et.al.(2020) **[58]**	2020	Multi-Attentional CNN	State-of-the-art deep fake detection	Fine-grained classification, multiple spatial attention heads, and textural feature enhancement
Hsu et.al.(2020) **[59]**	2020	Modified DenseNet	Effective discrimination in fake/real	Two-streamed network, pair wise learning, and classification layer
Nawaz et.al.(2023) **[60]**	2023	ResNet-Swish-Dense54	Robust deepfake detection	TailoredResNet-Swish-Dense54 framework, evaluation on
				Challenging datasets, adversarial attack testing
Yang et.al.(2021) **[61]**	2021	ResNet18 with Image Saliency	Precise detection of genuine and	Utilizes image saliency to uncover texture distinctions,

			manipulated facial images	improved guided filter for preprocessing
Gandhi et.al.(2020) **[62]**	2020	VGG, ResNet	Fortifying detection against adversarial attacks	Adversarial perturbations, Lipschitz regularization, Deep Image Prior (DIP), VGG and ResNet architectures
Nguyen et.al.(2019) **[63]**	2019	Capsule Network	Comparable performance to CNNs	Capsule-Forensics method, detection of various attacks, dynamic routing for agreement
Nguyen et.al.(2019) (Comprehensive) **[64]**	2019	Capsule Network	Versatility in various domains	Comprehensive investigation, random noise utilization during training
Marra et.al.(2018) **[65]**	2018	XceptionNet	Detecting modified images on social media	Identifying photos modified using GAN-based image-to-Image translation methods

XceptionNet is a convolutional neural network architecture designed to enhance image classification tasks by utilizing deeper and more intricate networks structures while retaining computational efficiency. By separating channel-wise and spatial convolutions, it reduce the number of parameters while maintaining the networks ability to capture complex patterns, resulting in better efficiency and performance compared to convolutional CNN architectures..

3. Conclusion and Future Work

Deepfake technology is evolving and has the potential to mislead many individuals. Although not every instance of deepfake content is malicious, detecting it is crucial because some of it poses serious risks. The primary aim of this study was to create a reliable and accurate technique for identifying deepfake images. Many other researchers have also been working diligently to detect deepfake content using a range of methods. The key contribution of this work is its success in obtaining remarkable results using CNN architecture. This study utilizes eight CNN architectures to identify deepfake images from a substantial dataset. The results have shown consistency and accuracy, with MTCNN excelling in several metrics, including accuracy, precision, F1-score and the area under the ROC curve. However, regarding recall, the custom model developed in the study slightly exceeded the performance of the MTCNN. The custom models, including DenseNet169, DenseNet201, VGG19, VGG16, ResNet50 and DenseNet121, also delivered images were analyzed for detection, yielding satisfactory outcomes.

This groundbreaking work is set to significantly impact society. With this technology, victims of deepfakes can swiftly ascertain whether images are genuine or fabricated. With our work, people will be able to stay vigilant by identifying deepfake images. Looking ahead, we plan to apply CNN algorithms to video deepfake datasets to assist more individuals.

Several experiments and tests remain to be conducted in future work. We intend to collect real data from our local community and use a convolutional neural network to classify deepfake images versus genuine ones. We plan to utilize more effective models to detect deepfake images, aiming to decrease crime both locally and globally. We believe our efforts will ultimately help reduce incidents of unwanted suicides

and blackmail in our society.

Reference:

1. Nguyen, T. T., Nguyen, Q. V. H., Nguyen, D. T., Nguyen, D. T., Huynh-The, T., Nahavandi, S & Nguyen, C. M. (2022). Deep learning for deepfakes creation and detection: A survey. Computer Vision and Image Understanding, 223, 103525.
2. Mirsky, Y., & Lee, W. (2021). The creation and detection of deepfakes: A survey. ACM Computing Surveys (CSUR), 54(1), 1-41.
3. Heidari, A., Jafari Navimipour, N., Dag, H., & Unal, M. (2023). Deepfake detection using deep learning methods: A systematic and comprehensive review. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, e1520.
4. Dolhansky, Brian & Bitton, Joanna & Pflaum, Ben & Lu, Jikuo & Howes, Russ & Wang, Menglin & Ferrer, Cristian. (2020). The DeepFake Detection Challenge Dataset.
5. Heidari, A., Jafari Navimipour, N., Dag, H., & Unal, M. (2023). Deepfake detection using deep learning methods: A systematic and comprehensive review. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, e1520.
6. Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., & 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).
7. Yogesh & Tanwar, Sudeep & Bhattacharya, Pronaya & Gupta, Rajesh & Alsuwian, Turki & Davidson, Inno & Mazibuko, ThokoZile. (2023). An Improved Dense CNN Architecture for Deepfake Image Detection. IEEE Access. PP. 1109/ACCESS.2023.3251417.
8. El-Gayar MM, Abouhawwash M, Askar SS, Sweidan S. A novel approach for detecting deep fake videos using graph neural network. Journal of Big Data. 2024 Feb 1.
9. Hanqing, Zhao & Wei, Tianyi & Zhou, Wenbo & Zhang, Weiming & Chen, Dongdong & Yu, Nenghai. (2021). Multi-attentional Deepfake Detection. 2185-2194. 10.1109/CVPR46437.2021.00222.**[10]**
11. Tran VN, Kwon SG, Lee SH, Le HS, Kwon KR. Generalization of forgery detection with meta deepfake detection model. IEEE Access. 2022 Dec 26.
12. Nirkin, Yuval & Wolf, Lior & Keller, Yosi & Hassner, Tal. (2020). DeepFake Detection Based on the Discrepancy Between the Face and its Context.
13. Sohail Ahmed Khan and Duc-Tien Dang-Nguyen. (2022). Hybrid Transformer Network for Deepfake Detection. https://doi.org/10.1145/3549555.3549588.
14. Ghai, A., Kumar, P., & Gupta, S. (2021). A deep-learning-based image forgery detection framework for controlling the spread of misinformation. Information Technology & People.
15. Sohail Ahmed Khan and Duc-Tien Dang-Nguyen. (2022). Hybrid Transformer Network for Deepfake Detection. https://doi.org/10.1145/3549555.3549588.
16. Zhao, H., Zhou, W., Chen, D., Wei, T., Zhang, W., & Yu, N. (2021). Multi-attentional deepfake detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 2185-2194).
17. Ghai, A., Kumar, P., & Gupta, S. (2021). A deep-learning-based image forgery detection framework for controlling the spread of misinformation. Information Technology & People.

18. 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.
19. 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.
20. 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.
21. 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.
22. 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. **[23]**
24. Sohail Ahmed Khan and Duc-Tien Dang-Nguyen. (2022). Hybrid Transformer Network for Deepfake Detection. https://doi.org/10.1145/3549555.3549588.
25. Munir, Kashif & Raza, Ali & Almutairi, Mubarak. (2022). A Novel Deep Learning Approach for Deepfake Image Detection. Applied Sciences. 12. 10.3390/app12199820.
26. A. Malik, M. Kuribayashi, S. M. Abdullahi and A. N. Khan, "DeepFake Detection for Human Face Images and Videos: A Survey," in IEEE Access, vol. 10, pp. 18757-18775, 2022, doi: 10.1109/ACCESS.2022.3151186.
27. 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. **[28]**
29. Agarwal, S.; Varshney, L.R. Limits of deepfake detection: A robust estimation viewpoint. arXiv 2019, arXiv:1905.03493.
30. 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.
31. 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. **[32]**
33. 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.
34. Kawa, P.; Syga, P. A note on deepfake detection with low-resources. arXiv 2020, arXiv:2006.05183.
35. 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.

36. 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].
37. 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.
38. Wubet, W.M. The deepfake challenges and deepfake video detection. Int. J. Innov. Technol. Explor. Eng. 2020, 9, 789–796. [CrossRef].
39. Pishori, A.; Rollins, B.; van Houten, N.; Chatwani, N.; Uraimov, O. Detecting deepfake videos: An analysis of three techniques. arXiv 2020, arXiv:2007.08517.
40. 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.
41. Vakhshiteh, F.; Ramachandra, R.; Nickabadi, A. Threat of adversarial attacks on face recognition: A comprehensive survey. arXiv 2020, arXiv:2007.11709.
42. Mirsky, Y.; Lee, W. The creation and detection of deepfakes: A survey. ACM Comput. Surv. (CSUR) 2021, 54, 1–41. [CrossRef]
43. 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]
44. 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.
45. Verdoliva, L. Media forensics and deepfakes: An overview. IEEE J. Sel. Top. Signal Process. 2020, 14, 910–932. [CrossRef].
46. 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].
47. 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.
48. 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.
49. 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.
50. Zhao, H., Zhou, W., Chen, D., Wei, T., Zhang, W., & Yu, N. (2021). Multi-attentional deepfake detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 2185-2194).
51. Gandhi, A., Jain, S. Adversarial Perturbations Fooldeepfake Detectors. arXiv preprint arXiv2003.10596, 2020. https://doi.org/10.1109/IJCNN48605.2020.9207034.

52. Nguyen, H. H., Yamagishi, J., & Echizen, I. (2019). Use of a capsule network to detect fake images and videos. arXiv preprint arXiv:1910.12467.
53. Ali, S. S., Ganapathi, I. I., Vu, N. S., Ali, S. D., Saxena, N., &Werghi, N. (2022). Image forgery detection using deep learning by recompressing images. Electronics, 11(3), 403.
54. Suganthi, S. T., Ayoobkhan, M. U. A., Bacanin, N., Venkatachalam, K., Štěpán, H., & Pavel, T. (2022). Deep learning model for deep fake face recognition and detection. PeerJ Computer Science, 8, e881.
55. Nataraj, L., Mohammed, T. M., Chandrasekaran, S., Flenner, A., Bappy, J. H., Roy-Chowdhury, A. K., & Manjunath, B. S. (2019). Detecting GAN generated fake images using co-occurrence matrices. arXiv preprint arXiv:1903.06836.
56. Guarnera, L., Giudice, O., Battiato, S. Deepfake Detection by Analyzing Convolutional Traces. Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, 666-667. ttps://doi. org/10.1109/CVPRW50498.2020.00341.
57. Ghai, A., Kumar, P., & Gupta, S. (2021). A deep-learning-based image forgery detection framework for controlling the spread of misinformation. Information Technology & People.
58. Zhao, H., Zhou, W., Chen, D., Wei, T., Zhang, W., & Yu, N. (2021). Multi-attentional deepfake detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 2185-2194).
59. Hsu, C. C., Zhuang, Y. X., & Lee, C. Y. (2020). Deep fake image detection based on pairwise learning. Applied Sciences, 10(1), 370.
60. Nawaz, M., Javed, A., & Irtaza, A. (2023). ResNet-Swish-Dense54: a deep learning approach for deepfakes detection. The Visual Computer, 39(12), 6323-6344.
61. Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., & 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).
62. Gandhi, A., Jain, S. Adversarial Perturbations Fooldeepfake Detectors. arXiv preprint arXiv2003.10596, 2020. https://doi.org/10.1109/IJCNN48605.2020.9207034.
63. Nguyen, H. H., Yamagishi, J., & Echizen, I. (2019). Use of a capsule network to detect fake images and videos. arXiv preprint arXiv:1910.12467. **[64]**