AN ENHANCED CONVOLUTIONAL NEURAL NETWORK FOR DETECTING DEEPFAKE VIDEOS

2023
PhD THESIS
COMPUTER ENGINEERING

Saadaldeen Rashid AHMED

Thesis Advisor
Assist. Prof. Dr. Emrullah SONUÇ

AN ENHANCED CONVOLUTIONAL NEURAL NETWORK FOR DETECTING DEEPFAKE VIDEOS

Saadaldeen Rashid AHMED

Thesis Advisor
Assist. Prof. Dr. Emrullah SONUÇ

T.C.
Karabuk University
Institute of Graduate Programs
Department of Computer Engineering

Prepared as
PhD Thesis

KARABUK
September 2023

I certify that in my opinion, the thesis submitted by Saadaldeen Rashid AHMED titled "AN ENHANCED CONVOLUTIONAL NEURAL NETWORK FOR DETECTING DEEPFAKE VIDEOS" is fully adequate in scope and quality as a thesis for the degree of PhD.

Assist. Prof. Dr. Emrullah SONUÇ .....

Thesis Advisor, Department of Computer Engineering

This thesis is accepted by the examining committee with a unanimous vote in the Department of Computer Engineering as a PhD thesis. September 13, 2023

Examining Committee Members (Institutions)	Signature
Chairman : Assoc. Prof. Dr. Rafet DURGUT (BANU)	.....
Member : Assoc. Prof. Dr. Adib HABBAL (KBU)	.....
Member : Assoc. Prof. Dr. Caner ÖZCAN (KBU)	.....
Member : Assist. Prof. Dr. Yusuf Yargı BAYDİLLİ (HU)	.....
Member : Assist. Prof. Dr. Emrullah SONUÇ (KBU)	.....

The degree of PhD by the thesis submitted is approved by the Administrative Board of the Institute of Graduate Programs, Karabuk University.

Assoc. Prof. Dr. Zeynep ÖZCAN .....

Director of the Institute of Graduate Programs

"I declare that all the information within this thesis has been gathered and presented by academic regulations and ethical principles and I have according to the requirements of these regulations and principles cited all those which do not originate in this work as well."

Saadaldeen Rashid AHMED

ABSTRACT

PhD Thesis

AN ENHANCED CONVOLUTIONAL NEURAL NETWORK FOR DETECTING DEEPFAKE VIDEOS

Saadaldeen Rashid AHMED

Karabük University
Institute of Graduate Programs
Department of Computer Engineering

Thesis Advisor:
Assist. Prof. Dr. Emrullah SONUÇ
September 2023, 95 pages

Deepfake detection is critical to address the proliferation of manipulated videos that can deceive and spread misinformation. Detecting deepfakes helps ensure the authenticity of visual content, protecting individuals, organizations, and society from potential harm, fraud, and misinformation. It safeguards trust in digital media and maintains the integrity of information in an era where video manipulation is increasingly sophisticated and accessible. Deepfake videos pose a significant threat to the integrity of visual content in the digital age. Detecting these manipulations is essential for safeguarding trust and authenticity. This research aims to enhance deepfake detection through the application of Rationale-Augmented Convolutional Neural Networks (RACNN) with Donald Trump Filter, addressing the urgent need to combat the proliferation of deceptive media and ensure the reliability of visual information. In our RACNN model, datasets play a crucial role in training. We have a total of 99,260 images, divided into two classes, with 70% for training. In addition,

there are 1,030 images for validation, which is 10% of the dataset, and 26,914 images for testing and fine-tuning, which is 20%. This setup helps ensure that our model can accurately distinguish between real and fake videos, contributing to the ongoing fight against deceptive digital content. In this thesis, we conducted an evaluation using two datasets: the Deepfake Detection Challenge (DFDC) and the FaceForensics++. The CNN approach remained consistent, resulting in minimal variation in computational cost between the two methods. When we applied the Donald Trump filter to Deepfake videos, we found that low computational cost was essential for making a faster connection based on facial associations. This large dataset has been replicated many times, making it ideal for accurate categorization and segmentation. In addition, the simple implementation of the CNN model allowed for seamless integration with a partitioning technique, resulting in impressive accuracy rates of 94.99% for the DFDC dataset and 93.99% for the FaceForensics++ dataset.

Key Word : Deepfake, video detection, segmentation, facial alignment, deep learning, reconstruction.

Science Cod : 92431

ÖZET

Doktora Tezi

DEEPFAKE VİDEOLARI TESPİT ETMEK İÇİN GELİŞTİRİLMİŞ EVRİŞİMLİ SİNİR AĞI

Saadaldeen Rashid AHMED

Karabük Üniversitesi
Lisansüstü Eğitim Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Tez Danışmanı:
Dr. Öğr. Üyesi Emrullah SONUÇ
Eylül 2023, 95 sayfa

Deepfake tespiti, yanıltıcı ve yanlış bilgileri yayan manipüle edilmiş videoların çoğalmasını engellemek için kritik öneme sahiptir. Deepfake'leri tespit etmek, görsel içeriğin orijinalliğini sağlamaya yardımcı olarak bireyleri, kuruluşları ve toplumu olası zararlardan, sahtekarlıktan ve yanlış bilgilerden korur. Dijital medyaya olan güveni korur ve video manipülasyonunun giderek daha karmaşık ve erişilebilir hale geldiği bir çağda bilginin bütünlüğünü korur. Deepfake videolar, dijital çağda görsel içeriğin bütünlüğüne yönelik önemli bir tehdit oluşturmaktadır. Bu manipülasyonların tespit edilmesi, güvenin ve özgünlüğün korunması açısından önemlidir. Bu araştırma, Donald Trump Filtresi ile Rasyonel-Artırılmış Evrişimli Sinir Ağlarının (Rationale-Augmented Convolutional Neural Network, RACNN) uygulanması yoluyla Deepfake tespitini geliştirmeyi, yanıltıcı medyanın çoğalmasıyla mücadele etme ve görsel bilgilerin güvenilirliğini sağlama konusundaki acil ihtiyacı karşılamayı amaçlamaktadır. RACNN modelimizde, veri kümeleri

eğitimde çok önemli bir rol oynamaktadır. Eğitim için %70'i iki sınıfa ayrılmış toplam 99.260 görüntümüz var. Ayrıca, veri kümesinin %10'unu oluşturan doğrulama için 1.030 görüntü ve %20'sini oluşturan test ve ince ayar için 26.914 görüntü bulunmaktadır. Bu kurulum, modelimizin gerçek ve sahte videoları doğru bir şekilde ayırt edebilmesini sağlayarak aldatıcı dijital içeriğe karşı devam eden mücadeleye katkıda bulunur. Bu tezde, iki veri kümesi kullanarak bir değerlendirme yaptık: Deepfake Detection Challenge (DFDC) ve FaceForensics++. CNN yaklaşımı tutarlı kalmış olup iki yöntem arasında hesaplama maliyetinde minimum değişiklikle sonuçlanmıştır. Donald Trump filtresini Deepfake videolarına uyguladığımızda, düşük hesaplama maliyetinin yüz ilişkilendirmelerine dayalı daha hızlı bir bağlantı kurmak için gerekli olduğu görülmüştür. Bu büyük veri kümesi birçok kez çoğaltılmıştır, bu da onu doğru kategorizasyon ve segmentasyon için ideal hale getirmektedir. Buna ek olarak, CNN modelinin basit uygulaması, bir bölümleme tekniğiyle sorunsuz entegrasyon sağlamıştır ve DFDC veri kümesi için %94,99 ve FaceForensics++ veri kümesi için %93,99 gibi etkileyici doğruluk oranlarıyla sonuçlanmıştır.

Anahtar Sözcükler : Deepfake, video algılama, segmentasyon, yüz hizalama, derin öğrenme, rekonstrüksiyon.

Bilim Kodu : 92431

ACKNOWLEDGMENT

First and foremost, I thank God Almighty for His blessing, which enabled me to complete this work of his grace. I would also like to thank the professors and Drs. who were credited with preparing this work and my supervisor Assist. Prof. Dr. Emrullah SONUÇ and jury members Assoc. Prof. Dr. Adil Deniz DURU and Assoc. Prof. Dr. Adib HABBAL, for their perseverance and supervision. I am grateful that they were my educational staff throughout the entire research period. Their counsel and assistance during the study have been helpful. Without their tireless aid, direction, and trust in my ability, completing this project would not have been feasible. I also thank them for opening my mind to a new universe of knowledge, chances, and experience, giving me a better understanding. In addition, I would like to express my gratitude to my parents and my brothers especially Mohammed Rashid Ahmed and my sisters, and my special thanks to my wife, who has supported me throughout my life and helped me earn my PhD. Finally, my wonderful friends A message of thanks and gratitude I send to you for always standing by my side; thank you so much for all.

CONTENTS

Page

Page

LIST OF FIGURES

Page

Page

LIST OF TABLES

Page

SYMBOLS AND ABBREVIATIONS INDEX

ABBREVIATIONS

AI	: Artificial Intelligence
AMTEN	: Adaptive Manipulation Traces Extraction Network
AUC	: Area Under the Curve
CGI	: Computer-Generated Imagery
CNN	: Convolutional Neural Network
DFDC	: Deepfakes Detection Challenge
DL	: Deep Learning
DNN	: Deep Neural Networks
GAN	: Generative Adversarial Network
KD	: Knowledge Distillation
LDPTOP	: Local Derivative Patterns on Three Orthogonal Planes
LRCN	: Long-Term Recurrent Convolutional Network
LSTM	: Long Short-Term Memory Network
ML	: Machine Learning
MLP	: Multi-Layer Perceptron
NT	: Neural Textures
PRNU	: Photo-Response Non-Uniformity
ProGAN	: Progressively Generative Adversarial Network
RACNN	: Rationale-Augmented Convolutional Neural Network
ReLU	: Rectified Linear Unit
RNN	: Recurrent Neural Network
SCNN	: Set Convolutional Neural Network
SPSL	: Spatial-Phase Shallow Learning
StyleGAN	: Style Generative Adversarial Network

PART 1

INTRODUCTION

In recent years, Deepfake technology has made significant advancements, allowing the creation of highly realistic and convincing fake videos of individuals. While Deepfake technology has potential applications in the fields of entertainment and media, it also poses a significant threat to the integrity of information and the security of individuals [1]. The ability to create fake videos of individuals can be used to spread misinformation, impersonate individuals, and even interfere in elections [2].

Deepfake refers to a technique that allows a user to superimpose their own face onto a video of an original person, making a recording that looks to perform or convey the same things as the real person. The face swap Deepfake variation fits within this category. Depending on the Artificial Intelligence (AI) employed to generate the content, Deepfakes could consist of lip-syncing or puppeteers. In Deepfake videos, the lip movements are coordinated with the music. Deepfakes of puppet masters are recordings of puppets mimicking the facial expressions, eye movements, and other actions of a human performer wearing a mask [3].

Deepfake can be generated using standard visual effects or Computer-Generated Imagery (CGI) tools. Despite this, some Deep Learning (DL) approaches like autoencoders and the Generative Adversarial Network (GAN) area of computer vision have been widely used. These models are used to produce new faces and body movements with similar expressions and behaviors to understand better how different facial expressions and body movements interact. Deepfake technologies occasionally require many images and video data to train the design for recognizing photorealistic images and videos. National personalities and celebrities, particularly candidates, are attractive targets for Deepfake assaults because of the abundance of online photographic and video information about them. Since then, the faces of

numerous celebrities have been substituted with pornographic females in movies generated utilizing the Deepfake technique [4].

While existing forgery detection systems perform well, the primary difficulty lies in their inability to generalize to newly emerging types of forgeries [5]. For example, a detector trained to identify face-swapping forgeries would struggle to perform accurately on facial reenactment forgeries. This makes forgery detection systems less practical since new types of forgeries are frequently appearing. Furthermore, supervised detection, which requires substantial training data on a specific forgery method, is not immediately capable of detecting newly emerging types of forgeries [6].

For instance, the use of Deepfake technology has potential disadvantages and ethical concerns such as the creation of false and misleading information, the potential for fraud or criminal activities, and the violation of individual privacy by creating fake content that can be used to defame or blackmail them. It is important to be aware of these issues and take appropriate measures to prevent and address them [7]. A video featuring Deepfake technology, depicting House Speaker Nancy Pelosi delivering a speech with noticeable speech impairment and displaying awkward behavior, has been widely shared on several social media platforms. The video sparked concerns about the potential for Deepfakes to be used to spread disinformation and interfere with political campaigns [8].

A Deepfake video of a Tom Cruise impersonator also went viral on TikTok. The video was created using Deepfake technology. It showed the impersonator performing a series of stunts and tricks that made it appear as though he was the real Tom Cruise. While the video was created for entertainment purposes, it raised concerns about the potential for Deepfakes to be used to impersonate public figures and spread disinformation [9].

The rationale-based models are used to analyze the text associated with the video, such as the title, description, and captions, to identify any inconsistencies or misleading information [9].

1.1. RESEARCH PROBLEM

The majority of Deepfake detection techniques rely on the utilization of Convolutional Neural Networks (CNNs) to extract frame-level information from an image [10]. Temporal features, as a category of detection methodology, are limited in their ability to capture concealed elements within a video's temporal progression. Recent studies have investigated these temporal characteristics, revealing their superiority compared to techniques based on individual frames [11]. There is a lack of studies that have used hand-crafted facial features to enhance classification performance, despite the discovery made by [12] that such traits may offer additional information to models. Also, we found that Deepfake datasets are highly oversampled, causing models to become easily overfitted. The datasets are created using a small set of real faces to generate multiple fake samples. When trained on these datasets, models tend to memorize the actors' faces and labels instead of learning fake features [13].

The solution to this problem, we develop a Rationale-Augmented Convolutional Neural Networks (RACNN) based a novel methodology for deepfake face reconstruction with Donald Trump filter.

1.2. AIM OF THE STUDY

This thesis first provides an overview of the current state-of-the-art Deepfake detection methods and their limitations. We then present the proposed RACNN method with Donald Trump Filter and its implementation. When applying the Donald Trump filter to the Deepfake video, we found that a low computational cost was necessary to establish a faster link based on the association between the faces. The performance of the proposed method is evaluated on two datasets and compared to the state-of-the-art methods. Therefore, it is crucial to develop methods for detecting Deepfakes to maintain the trust and integrity of information.

i. Providing an overview of previous works and achievements on reconstruction of different type of faces.

ii. Developing a RACNN based novel methodology for deepfake face reconstruction with Donald Trump Filter.
iii. Applying RACNN model to two different datasets to perform reconstruction with less data loss and more accuracy.
iv. Evaluating and comparing the proposed RACNN model with other studies in the literature.

Overall, rationale augmentation in CNNs enhances their transparency and the ability to provide meaningful justifications for their predictions, making them more useful and trustworthy in various applications.

1.3. THESIS STRUCTURE

This thesis is structured into five chapters, each focusing on specific aspects of the research on Deepfake detection using DL.

The first chapter, “Introduction”, provides an overview of the problem statement, the aim of the research, and the potential uses of DL.

The second chapter, “Literature Review”, presents a thorough analysis of the current research that has employed DL and Machine Learning (ML) for Deepfake detection.

The third chapter focuses on “Types of Digital Face Manipulations” and provides a thorough examination of numerous digital face manipulation techniques, offering detailed explanations and analysis for each type.

The fourth chapter, “Methodology”, details the implementation procedures of the classification method used in the study.

In the fifth chapter, “Results & Discussion”, the comprehensive results and their discussion are presented, showcasing the thesis’s overall findings.

Finally, the sixth chapter, “Conclusion”, summarizes the study's contributions, assesses the interpretation of the findings, and discusses potential future development. By dividing the thesis report into these six chapters, the reader can gain a comprehensive understanding of the research and its contributions to the field.

PART 2

LITERATURE REVIEW

Detecting deepfakes is an important area of research and development to counter the potential misuse of this technology. There are several methods and techniques used for deepfake detection, facial anomaly detection, and body anomaly detection. Deepfakes often have subtle anomalies in facial features or body movements that may not be perfectly aligned. These can be detected using computer vision techniques to analyze inconsistencies in facial expressions, eye blinks, lip sync, or strange artifacts.

There are three main categories of algorithms used to modify visual attributes within video content. The simplest approach is the graphics-based technique [14]. This is commonly used in basic smartphone applications such as Snapchat. The second type of algorithm uses a latent feature space to identify and differentiate individuals based on their facial features. These techniques are more complex and require the process of training an auto-encoder on a specific set of targets [16]. Finally, it should be noted that other GAN-based methods are available [16]. However, it is important to recognize that these are predominantly used to generate static rather than dynamic visual content. A variation of the following three core principles is used in several widely used open-source programs. These implementations may have graphical user interfaces that are designed to be user-friendly. This improves accessibility to a wider range of users, including the public. Because of their user-friendly interfaces and wide accessibility, DeepFaceLab [14], Deepfake [15] and FaceSwap [16] are widely used tools for generating a significant amount of deepfake content distributed on the Internet as shown in Figures 2.1 and 2.2.

Figure 2.1. Example of face swapping with a graphics-based method [16].

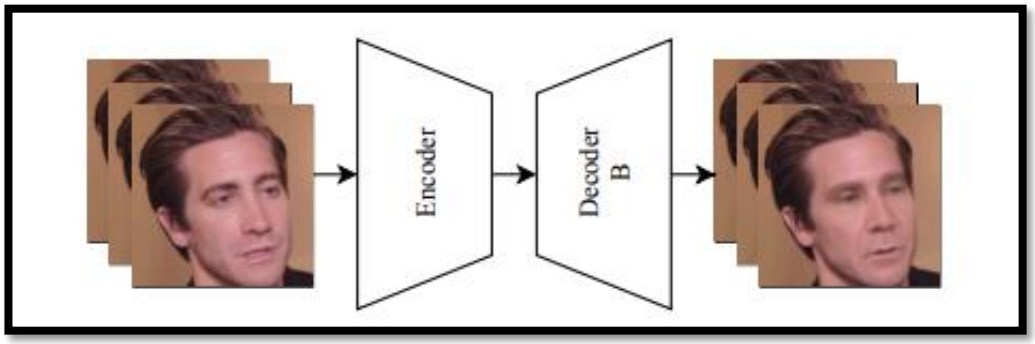

Figure 2.2. The generation process of the encoder-decoder method [16].

In 2015, Deepfake demonstrated a satisfied accuracy in the widely recognized Recurrent Neural Network (RNN) benchmark, achieving an 87.80% over the subsequent three years. This realization gave rise to a plethora of data-efficient topologies. The review also refers to various DL architectures [17], such as specific variations of the Inception architecture (v1 and v2) and covers the evolution of facial reconstruction and various approaches used to accomplish the process.

Researchers have introduced a Deep Neural Network (DNN) facial reconstruction method, utilizing facial landmarks like eyes, nose, mouth corners, and face center to generate a 160-dimensional Deepfake vector. This vector aims to predict over 10,000 unique individuals within the resulting matched IDs [18]. The design predicts that n > 10,000 different individuals will be represented in the final count of matched IDs. Figure 2.3 presents an architectural diagram for CNN with a deeply hidden identity.

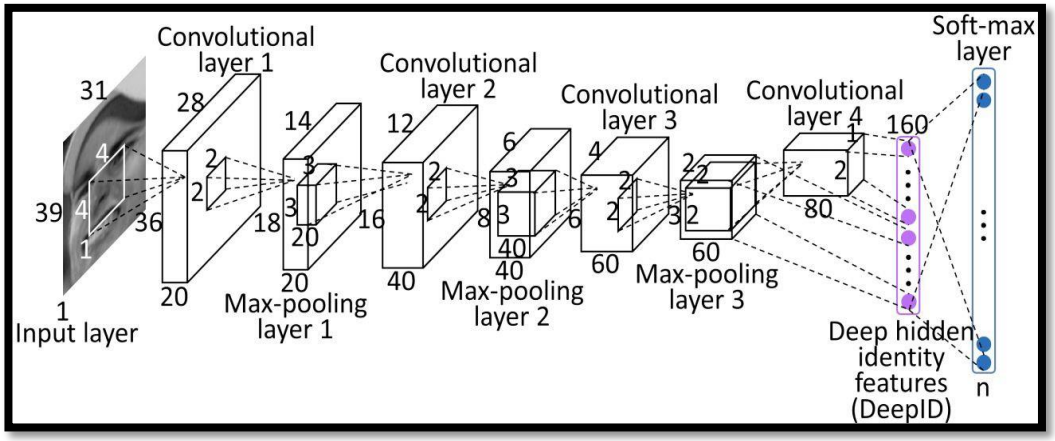

Figure 2.3. An architectural diagram for the CNN with a deeply hidden identity [19].

DeepFake was followed by the development of DeepFake2, DeepFake2+, and DeepFake3, with each iteration introducing modifications to the underlying concept. These models reportedly used a variety of facial landmark detectors along with different patch or landmark selection criteria [19].

The distinctiveness of individuals is mostly determined by the unique characteristics of the human face. Regrettably, the proliferation of face-synthesizing technology is emerging as a plausible threat to societal well-being. Numerous sets of rules grounded in DL knowledge exist, enabling the substitution of real human faces with synthetic counterparts. Deepfake is an emerging field within the realm of AI that pertains to the process of overlaying the facial features of one individual onto the visage of another. GANs have the capability to generate Deepfake images of superior quality, as evidenced by recent research [20].

The rapid dissemination of Deepfake material has been facilitated by the proliferation of mobile devices and public websites [21]. The occurrence of pixel collapse, a phenomenon characterized by the appearance of unexpected visual irregularities in the tonal quality or facial appearance of photographs, was the initial factor that made Deepfake images perceptible to human observers. Deepfakes can be created using both auditory and visual data. In recent years, significant progress in technology has led to the development of Deepfakes, which have reached a level of visual fidelity that makes them nearly indistinguishable from real photographs [22].

As a result of this phenomenon, a wide variety of issues affect a significant number of individuals on a global scale.

Generally, Deepfake offers certain positive benefits despite the numerous disadvantages accompanying it. For instance, Deepfake is advantageous to the fashion and e-commerce industries since it expedites consumer purchasing. The Deepfake technology also assists the music industry by offering artificial voices to musicians incapable of easily naming their work. Additionally, Deepfake enables filmmakers to replicate or reuse the special effects from various memorable sequences. Moreover, advanced communication and Deepfake technologies may assist Alzheimer's patients in retaining more memories. Detecting abnormalities in X-ray images using GANs is also the focus of ongoing research [23]. The witnesses need many photos, videos, or audio samples to create convincing fakes using the Deepfake approach.

Unfortunately, Deepfake technology poses a significant threat to many especially public figures. Regrettably, being the center of attention comes with significant disadvantages. The vast quantity of publicly accessible media featuring, among others, celebrities, athletes, and politicians makes them easy targets for Deepfakes. Deepfake technology is mostly used to make mock others. For political, sexual, or comedic objectives, one can exploit friends' voices and photographs without consent. Famous faces of pornographic models can also be found online [24]. Today, fake content is easy to make [25].

Cyberbullying is a growing problem among young people, with tragic consequences such as suicide. A video showing the former president of the United States, Barack Obama, allegedly expressing statements he has never spoken is rapidly circulating online. Deepfakes have been utilized to change Joe Biden's tongue-out appearance in the 2020 presidential campaign video, it is important to acknowledge that social media platforms can pose substantial risks to both people and society at large. These risks encompass the dissemination of misinformation, breaches of personal privacy, damage to one's reputation, fraudulent impersonation, and manipulation of public sentiment [26]. Many women outside of Asia and the United States are affected by

Deepfakes technologies. When Deepfakes are utilized on social media, deceptive material can spread far more rapidly and have far-reaching cultural effects [27]. However, due to the tremendous harm inflicted upon individuals and businesses, it is imperative that sophisticated forgeries be given substantial consideration in the design process. For this reason, scientists have worked diligently to uncover the phenomenon of Deepfakes to protect the public from instances of defamation, fraudulent schemes, misleading information, and vulnerabilities.

The exposure of Deepfakes has the potential to significantly decrease global crime rates. The focus of the scientific community has been directed towards the validation procedure pertaining to these statements [28]. Unfortunately, only a limited number of prominent multinational corporations have implemented measures in reaction to this prevailing pattern. Leading technology companies such as Google, Facebook, and Microsoft have made a comprehensive dataset of fabricated movies available to the research community [29]. This valuable resource enables researchers to develop novel algorithms and methodologies for effectively identifying and mitigating the spread of Deepfake content.

Overall, DeepFake models performed admirably with the GAN, achieving an accuracy of 89.53% [30]. However, when the number of classes expanded, difficulties emerged [30]. As a result of a considerable increase in types and a drop in the number of face photos associated with each class, the authors determined that “the classifier outputs were diverse and unreliable, and hence cannot be used as features” [31]. The CNN model was trained to predict the best class from a given set of IDs, but it lacked a common approach for formalizing visual traits among otherwise comparable faces. Face verification using DeepFake is expected to be significantly less accurate for those outside the 10,000 classes [32-34].

FaceNet's face alignment step was one of its defining aspects; it was utilized to enhance the model. Figure 2.4 illustrates how challenging it means to execute this technology because training the model requires a large dataset, which only Facebook possesses. The model's final architecture was determined to output a 4096-

dimensional representation vector, trained with cross-entropy loss, to predict the probability of classes' association [35].

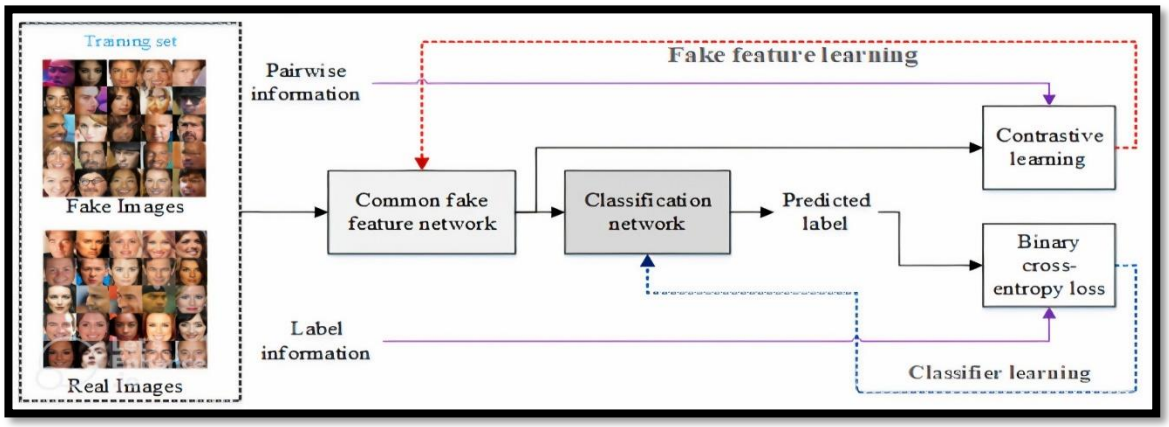

Figure 2.4. Different orientations of faces are being evaluated for the Deepfake network [36].

FaceNet introduced the FaceForensics++ dataset, a significant contribution to the field of Deepfake detection, along with a comprehensive benchmarking framework. This dataset was instrumental in shedding light on the daunting challenges of detecting increasingly realistic deepfake content [37].

A method for generating realistic video portraits by transferring facial expressions from a source to a target actor. It highlights the challenges of distinguishing between real and generated videos and the need for effective detection mechanisms [38]. Explore the task of face anti-spoofing, which is related to Deepfake detection. The paper presents a framework that leverages auxiliary supervision signals to improve the detection accuracy of deep learning models [39].

The method for detecting more general face forgery techniques, including Deepfakes and traditional image editing. The authors introduce the "Face X-ray" concept and use a deep neural network for detection and explores the use of lip movements to detect Deepfake videos [40]. The authors then propose a convolutional neural network-based approach that analyzes lip synchronization and motion for detection [41].

Following the conversion of the labeled matrix into an image for face detection and reconstruction [42], DL techniques are employed to represent the resulting output visually. Before applying DL algorithms to complement the image, several pre-processing approaches are employed to enhance the region of interest intended to be extracted. This includes using the Long Short-Term Memory Networks (LSTM) approach and a morphological erosion technique using a grain size distribution to enhance contrast in the facial region. The proposed approach consistently detects regions of interest within photos that exhibit sufficient contrast. This enables the DL algorithm to extract the drainage basins of the Deepfake network successfully [43].

DL-based techniques aid in face extraction from media through blob detection and fake face segmentation [44, 45]. The Deepfake face segmentation approach is robust in varying illumination conditions, extracting even small sections, including the secret area of interest [46]. Scholars have used Deepfake feature analysis to track desired attributes and extract data [47]. Face reconstruction and Deepfake networks pinpoint areas for ellipse-based extraction using these features [48].

For face analysis, recording features within a specified area range and threshold is crucial. A threshold of 12 and an area range of 200 to 5000 were used; lower thresholds identified more regions, but higher thresholds missed expected regions [49]. Sometimes a threshold of 12 failed to recover multiple faces due to the wide region range or slightly low threshold [50]. Thus, selecting an appropriate threshold and range is vital for trustworthy DL algorithm results [51].

Photographs of a face mask are taken, then the images are reconstructed using the information recorded by the auto-encoder. Switching between the source and destination images' appearances necessitates using two codecs, and the encoder parameters for both pairs of codecs are shared in this instance. Each group is being utilized to expand its photography knowledge. In both sets, the encoder networks are identical. This encoder-decoder architecture is used by Deepfakes Faker [52], Deep Facet “TensorFlow-based Deep-fakes” [53], and DF [54]. An advanced form of Deepfakes uses an innovative adaptive GAN for face transformations. This GAN

addresses the shortcomings in encoder and decoder design seen in VGG-face, aiming to overcome adversarial constraints, and missed perceptual opportunities [55-57].

FaceNet's implementation of a CNN with several tasks is intended to improve face recognition and orientation constancy. Cycle GAN is a helpful method for constructing generative networks [58]. The impact of Deepfakes on the safety, privacy, and authenticity of democratic republics has been a growing concern [59]. To mitigate these risks, a constant surveillance system has been implemented to ensure that Deepfakes are detected and addressed promptly. Recent advancements in DL have enabled automatic Deepfake detection [60]. To overcome this issue, Korshunov and Marcel [61] implemented the freely available Face swap-GAN [62] technique to generate a (620-record) GAN-shaped Deepfake dataset. Low-budget, high-rating Deepfake movies have been produced using the VidTIMIT dataset [63] to imitate gaze blink, brim movement, and facial scrub expressions [64]. The findings of experiments indicate that typical facial recognition systems cannot identify deeply faked faces. Due to the improved effectiveness of DL techniques such as CNN and GAN in strengthening image readability, face representation, and illumination, forensics models have gotten more complex [65].

The literature on the detection of identity-swapped movies has grown over time [66]. In identity-switching videos [67], the facial appearance of one individual is replaced with that of another. Most of these algorithms use an auto-encoder architecture to reconstruct the facial features of the desired individual [68]. The auto-encoder consists of a system that includes both an encoder and a decoder. The process of encoding the input face yields a latent vector that serves as a representation of the information conveyed by the face. The encoder uses the same set of weights, ensuring that it learns properties that are independent of the identity being encoded [69]. However, the decoder uses the latent vector to reconstruct the facial features of the individual under consideration [70]. This approach is used by several deepfake generation systems, including DFaker and DeepFaceLab. The architecture of the face swap GAN auto-encoder has been further improved by incorporating adversarial and perceptual loss into the algorithm. The integration of adversarial and perceptual loss was implemented in the synthesis procedure to minimize the occurrence of visual

artifacts [66]. As a result, distinguishing real videos from deepfakes is an increasingly difficult task for detection models.

Various methods can be used to identify videos in which the facial features of two people have been swapped during the editing process. Over time, many innovative deepfake detection methods have been documented in the academic literature. The compilation of studies included in Appendix A provides a comprehensive overview of the most relevant research conducted on the topic of identifying instances where recordings have undergone identity switching. Most of this research has been documented in the review article [66,67]. The table has been revised to include the results of more recent studies that have produced more recent results. Deepfake movies often show visual distortions resulting from the switching of two faces. As a result, early efforts in this area relied heavily on manually created attributes derived from these distortions [71]. To identify individuals engaged in deepfake impersonation, a technique based on the calculation of 3D distance vectors between mouth landmarks, head rotations and facial activity units has been introduced [72]. According to the authors, it is claimed that an individual's facial expressions and bodily gestures show distinctiveness during the act of verbal communication [72]. It is intriguing that the researchers have linked the mouth to at least one of the top five distinctiveness attributes. Ultimately, the most optimal model achieved an Area Under the Curve (AUC) of 96.3% on their proprietary dataset. In addition, a novel technique was developed to detect deepfakes based on their visual irregularities, including inaccuracies in geometry and inaccurate assessments of lighting conditions, particularly around the eyes and teeth. The Multilayer Perceptron (MLP) proved to be the most successful design, achieving an AUC of 85.1% when evaluated on a proprietary database [73]. When the performance of this model on the Celeb-DF dataset yielded an AUC of 55.0% [69], this result can still be considered a commendable performance.

Another approach involves the use of artifacts introduced during the synthesis process [74]. The researchers postulated that the use of only peripheral facial features would result in different 3D head position estimates compared to the use of the full set of landmarks. The performance of their technique achieved an AUC of 89.0% on

the UADFV dataset, but only an AUC of 54.6% when examined on the Celeb-DF dataset [69]. This study aims to demonstrate of the ability of a pre-trained ResNet50 model to discriminate between authentic videos containing facial distortion anomalies and deepfake videos [75]. The approach achieved an AUC of 64.4% when evaluated on the Celeb-DF dataset [75]. The analysis focuses on investigating strategies that rely on inconsistencies in eye blink patterns to detect deepfake generators, despite their limited ability to produce convincing results [76,77].

Image distribution is facilitated using a Long-Term Recurrent Convolutional Network (LRCN) trained on a dataset specifically curated for this task. The performance of this LRCN model is evaluated using the AUC metric, resulting in an impressive score of 99.0%. The Eye Aspect Ratio for each image was derived by considering demographic factors such as age, gender, level of physical activity and time of day. These values were then compared with established normative data on blinking behavior. The data were obtained from a separate repository. A distance measure was then used to authenticate the film using a proprietary data set, resulting in an accuracy rate of 87.5% [78]. There is an optimistic outlook for a future where visual artifacts will be less important as defining elements. The implementation includes detection techniques [72] that use DL algorithms capable of autonomously identifying discriminative features. To investigate mesoscopic features in photographs, we performed an analysis on two condensed deep learning networks. The best-performing model on the Celeb-DF dataset achieved an AUC value of 54.8%. This finding is of great interest as it highlights the importance of the oral and eye regions in identifying deepfakes [79].

The use of a two-channel network to detect altered facial images. One stream uses data obtained using a CNN to assess whether a facial image has been altered [80]. The alternative approach uses a patch triplet stream to retrieve information about hidden sources of noise, such as camera characteristics. Finally, the results from both streams are combined to determine the authenticity or manipulation of a given face [81]. Remarkably, a comprehensive investigation revealed that this particular methodology exhibited a high level of security when tested against the unobserved identity swap dataset [69]. Furthermore, a multi-task learning network has been

developed with the aim of detecting deepfake videos [66]. The approach described in this study is based on an auto-encoder architecture [66]. It proposes to facilitate the exchange of information between two tasks to improve the performance of both models. The effectiveness of the network was demonstrated on the FaceForensics++ dataset [82], although it showed shortcomings when applied to the Celeb-DF dataset [69]. Subsequently, researchers have demonstrated the ability of capsule networks to identify deceptive images and videos under various forms of attack [83]. Capsule networks are used to identify fake photos and videos by extracting latent information from a specific subset of the VGG-19 network [84]. The generalizability of the method was shown to be limited when evaluated on the Celeb-DF dataset [69]. In their study, researchers investigated the effectiveness of using specific facial regions as input to a pre-trained Xception network to identify the most discriminative region for deepfake detection [67].

Using the eye region as input gave the best results when applied to the Celeb-DF dataset. However, the highest level of performance, as indicated by an AUC of 83.6%, was achieved when the entire face was used as input. Finally, the authors emphasised the advantages of using second-generation deepfake datasets such as Celeb-DF and DFDC over first-generation datasets such as FaceForensics++ and UADFV [67]. Most detection algorithms rely on frame-level data. However, a time-aware technique has been developed to identify movies with identity swaps [85]. The researchers used a pre-trained InceptionV3 network to extract frame-level features. In addition, they used an LSTM network to detect anomalies across frames. The approach was evaluated on a proprietary dataset and achieved an accuracy of 97.1% based on a limited sample size of 40 photographs. It is worth noting that the majority of current deepfake detection techniques do not incorporate frame discrepancy learning [86].

In another study, researcher evaluated the effectiveness of five temporal action detection techniques in detecting deepfake videos. The strategy they employed shows superior performance in terms of AUC on the Celeb-DF dataset [86], surpassing the performance of existing frame-based detection approaches that are considered state-of-the-art. In addition, an evaluation of three state-of-the-art 3D CNN techniques was

performed, which showed significant potential in the field of action detection [87]. The best-performing model, ID3, achieved a accuracy of 95.1% on the FaceForensics++ dataset, which included identity switching [88]. Like the work, proposed a methodology for distinguishing authentic movies from deepfakes [89].

Hyper-realistic synthetic environments have been developed thanks to the widespread availability of cutting-edge computer vision and machine learning technology in recent years videos, commonly known as deepfakes, which have found applications spanning various domains. These developments, while enabling numerous positive use cases, have simultaneously fueled concerns about the urgent need for reliable deepfake detection systems [90]. underscores the limitations of traditional handcrafted approaches and sets a strong baseline for facial manipulation detection using modern deep learning architectures. Addressing the pressing issue of deepfake videos [91]. introduces a temporal-aware detection pipeline utilizing convolutional and recur- rent neural networks (CNN and RNN). With impressive accuracy exceeding 97%, this research demonstrates the capability to accurately identify manipulated video fragments, offering a potential solution to combat the proliferation of deepfake videos. Considering the increasing danger posed by computer-generated music and movies [92]. introduces a biometric forensic method for identifying" deep fakes" created by swapping faces. Rather of relying just on face recognition technology, this approach integrates temporal, behavioral biometrics based on facial emotions and head movements. With a remarkable detection accuracy exceeding 90%, this technique serves as a robust defense against deceptive media manipulations. To combat the deepfake phenomenon [93]. introduces a Capsule Network-based approach that exhibits remarkable versatility in detecting various forms of attacks, achieving an accuracy rate surpassing 90%. This method stands as a robust solution to safeguard against digital image and video manipulations, highlighting the potential of capsule networks in digital forensics.
In the realm of multimedia forensics [94]. tackles the problem of identifying fake videos on social media. Through exhaustive performance evaluations, this research reveals the importance of fine-tuning operations and presents valuable insights into identifying specific manipulation techniques employed in real-world scenarios involving shared manipulated media [95]. addresses the pressing challenge of

detecting manipulated faces in video sequences. The authors propose an ensemble approach that combines various Convolutional Neural Network (CNN) models, achieving promising detection accuracy on extensive video datasets, with specific accuracy figures available. This work highlights the importance of advanced detection methods in the age of manipulated video content [96]. introduces the VideoForensicsHQ benchmark dataset, challenging existing forgery detectors in detecting imperceptible manipulations. By combining geographical and temporal information, the authors create a new class of detectors, achieving an impressive detection accuracy range of 99.25% to 99.38%. Their detectors provide a robust defense against advanced manipulation techniques [97]. provides a novel method for identifying videotaped facial manipulation. The authors introduce the Set Convolutional Neural Network (SCNN) framework and evaluate its performance on various datasets, achieving state-of-the-art results. While specific accuracy figures are not mentioned in the summary, they can be found in the paper [98]. addresses the challenge of identifying tampered faces in video sequences using recurrent convolutional models. On video-based facial modification benchmarks, the authors' results are state-of-the-art, specifically detecting Deepfake, Face2Face, and FaceSwap manipulations, with specific accuracy figures available in the paper. Their approach surpasses the previous state-of-the-art by up to 4.55% in accuracy on the FaceForen- sics++ dataset, demonstrating its effectiveness in addressing the growing concern of misinformation in online content, particularly in the domain of video-based face manipulation detection. This work highlights the effectiveness of their approach in detecting misinformation in online video content [99]. Using the FaceForensics++ dataset, we will look at the performance of three innovative 3D CNN techniques for identifying doctored films. The study reveals promising results, with these 3D CNN models exhibiting strong detection accuracy across various manipulation techniques. Specifically, 3D ResNet achieved accuracy scores of 91.81% for Deepfake (DF), 89.6% for Face2Face (F2F), 88.75% for FaceSwap (FS), and 73.5% for NeuralTextures (NT). Similarly, 3D ResNeXt demonstrated competitive performance with accuracy rates of 93.36% (DF), 86.06% (F2F), 92.5% (FS), and 80.5% (NT). I3D exhibited impressive detection capabilities, with accuracy scores of 95.13% (DF), 90.27% (F2F), 92.25% (FS), and 80.5% (NT).Nguyen et al. [100]. In order to simultaneously detect image and video modifications and

identify the modified regions, this study proposes a convolutional neural network (CNN) that employs multi-task learning. The suggested technique was tuned on the FaceForensics and FaceForensics++ databases, where it attained an accuracy of 83.71% for classification and 93.01% for segmentation. demostrating its adaptability even to previously unseen manipulation techniques [101]. introduces a comprehensive benchmark for facial manipulation detection based on techniques such as DeepFakes, Face2Face, FaceSwap, and NeuralTextures. Trained forgery detectors, aided by domain-specific knowledge, achieve remarkable accuracy on the benchmark. Notably, the XceptionNet model achieves accuracy scores of 96.36% (DeepFakes), 86.86% (Face2Face), 90.29% (FaceSwap), 80.67% (NeuralTextures), and 52.40% (pristine images), resulting in an overall total accuracy of 70.10% [102]. presents a modified AlexNet-based model for detecting fake videos on publicly avail- able datasets. The model achieves high accuracy, with a 98.73% accuracy rate on the UADFV dataset and a 98.85% on the Celeb-DF dataset, how accurate you are. Using data from FaceForensics++, it achieves an accuracy rate of 87.49% for multiclass classification and binary classification [103]. The proposed method utilizes deep learning with compact network architectures for detecting face tampering in videos. For Deepfake videos, it detects more than 98%, and for Face2Face videos, it detects more than 95%.To better detect manipulation traces in facial photographs [104]. introduce the Adaptive Manipulation Traces Extraction Network (AMTEN). By combining a convolutional neural network (CNN) with a fake face detector (AMTEN- net), it is possible to obtain an excellent average accuracy of up to 98.52 percent when identifying false face photos created using different modification techniques. Further- more, AMTEN net keeps an average accuracy of 95.17 percent even when presented with face photos that have undergone unidentified post-processing processes [105]. addresses the growing threat of Deepfake videos by introducing an innovative forensic technique that uses optical flow fields to detect differences between fake and original video sequences. Initial experiments on the FaceForensics++ dataset show promising results, suggesting the potential of utilizing temporal dissimilarities for deepfake video identification [106]. presents a new way to find fake video sequences by looking at the dynamics of spatiotemporal texture, with a focus on the study of multiple temporal parts together. Using Local Derivative Patterns on Three Orthogonal Planes (LDP-TOP) as feature

markers, the suggested method is very good at telling the difference between real and fake video sequences. Lin- ear Support Vector Machines (SVMs) are used to classify data, and their success is comparable to that of deep models for finding fake material [107]. In response to the emerging threat of Deepfakes, this paper introduces OC-FakeDect, a one-class anomaly detection approach for Deepfake detection.OC-FakeDect only trains on pictures of real faces and thinks of Deepfakes as oddities, achieving a remarkable accuracy of 97.5% on the NeuralTextures dataset from the FaceForensics++ bench- mark, all without using any fake images for training. This one-class-based approach demonstrates its potential for robust Deepfake detection [108]. proposes FRe- TAL, an approach to transfer learning for deepfake detection that uses Representation Learning (ReL) and Knowledge Distillation (KD) to reduce the risk of catastrophic forgetting. Results from the FaceForensics++ dataset show that FReTAL excels in domain adaptation challenges, with experimental results showing accuracy of up to 86.97% on low-quality deepfakes [109] response to the increasing concerns posed by facial manipulation technologies, presents a new face forgery detection method based on frequency-aware discriminative feature learning (FDFL). This method produces state-of-the-art results on three variants of the FF++ dataset by combining a single-center loss (SCL) with an adaptive frequency feature generating module, addressing the need for more discriminative and data-driven techniques in this critical domain of computer vision [110] introduces the A Facial Imitation Detection Method Using Spatial-Phase Shallow Learning (SPSL), leveraging both spatial and frequency information, particularly emphasizing the importance of phase spectrum in detecting up-sampling artifacts. SPSL achieves state-of-the-art performance on cross-dataset evaluations and multi-class classification, offering a com- prehensive solution considering the security issues posed by sophisticated face forgery methods [111]. introduces a novel approach using capsule networks for detecting various forms of image and video forgeries, including computer-generated content and replay attacks. The proposed method outperforms other state-of-the-art algorithms and achieves great accuracy, with a 99.13% success rate on the FaceForensics++ dataset and a 96.75% success rate in distinguishing CGIs from PIs. This research demonstrates how capsule networks may be used in other areas of computer vision and forgery detection.

The model used in their study uses a bidirectional Gated Recurrent Unit network in conjunction with DenseNet for feature extraction. When evaluated on the FaceForensics++ dataset, the model achieved an AUC score of 99.4%. Different strategies can be used to detect and discriminate different video frame dynamics. A comparative analysis was conducted to identify deepfake movies by evaluating several DL algorithms [85]. One solution used a triplet loss architecture and metric learning techniques. This technique is used to distinguish different from similar features within the feature space. The approach was evaluated using the FaceForensics++ dataset and achieved an AUC of 92.9% using a limited number of 25 frames per video. The researchers used a different methodology, training an Xception network in its entirety using face data. As a result, they achieved the highest AUC (99.2%) on the Celeb-DF dataset. Additional information can be beneficial for classification models, and research has shown that manually generated features can contribute to this improvement [68]. Furthermore, the identification of facial emotions and human behavior has been improved by using both hand-crafted and deep features [112]. Researchers have integrated the dynamic attributes derived from deep learning with hand-crafted features to develop a method that effectively identifies facial expressions with high accuracy. DNNs may not adequately account for low-level geometric or appearance-based information, which can be effectively represented by hand-crafted features [112]

Table 2.1. Comparison table for related work to shown Main Findings Contributions, Limitations Challenges, algorithm used, strength and weaknesses.

Ref	**Main Findings Contributions**	**Limitations Challenges**	**Algorithm Used**	**Strengths**	**Weaknesses**
[38]	Addresses the generation of realistic video portraits and the need for deepfake detection	Challenges in distinguishing between real and generated videos	Facial expression transfer, deepfake detection	Addresses the challenge of generating realistic videos	Limited information on the specific algorithms used
[39]	Presents a framework	Challenges in face anti-	Deep learning,	Enhances deepfake	May require labeled data

	for improving deepfake detection accuracy using auxiliary supervision signals	spoofing detection	auxiliary supervision signals	detection accuracy	for auxiliary supervision
[40]	Introduces the "Face X-ray" concept for detecting face forgery techniques including Deepfakes	Detection of various face forgery techniques	Deep neural network, "Face X-ray" concept	Addresses a range of face forgery techniques	
[41]	Proposes a convolutional neural network-based approach for deepfake detection using lip synchronization	Detecting deepfakes using lip movements	Convolutional neural network, lip synchronization	Focuses on a specific aspect of deepfake detection	May not cover all types of deepfake manipulations
[42]	Converts labeled matrices into images for face detection and reconstruction using DL techniques	Challenges in face detection and reconstruction	Deep learning techniques, face detection	Offers a novel approach to face detection and reconstruction	Limited information on the specific algorithms used
[43]	Uses DL algorithms to successfully extract the drainage basins of the Deepfake network	Challenges in extracting drainage basins	DL algorithms, drainage basins extraction	Provides insights into Deepfake network analysis	Limited information on the specific DL techniques used
[44]	Utilizes DL techniques for face extraction	Challenges in blob detection for face	DL techniques, blob detection	Effective in extracting faces from media	May not cover all types of media or

	from media using blob detection	extraction			manipulation types
[45]	Develops a robust Deepfake face segmentation approach for varying illumination conditions	Challenges in robust face segmentation for Deepfakes	Deepfake face segmentation, robustness	Addresses the issue of varying illumination conditions	Limited information on the specific segmentation method
[46]	Uses Deepfake feature analysis to track desired attributes and extract data	Challenges in tracking attributes in Deepfakes	Deepfake feature analysis, attribute tracking	Provides insights into tracking attributes in Deepfakes	Limited information on the specific feature analysis
[47]	Pinpoints areas for ellipse-based extraction using Deepfake networks	Challenges in ellipse-based extraction	Face reconstruction, ellipse-based extraction	Offers an approach to extract regions of interest	Limited information on the specific extraction method
[48]	Discusses the importance of selecting an appropriate threshold and range for face analysis	Challenges in threshold selection for face analysis	Threshold selection, face analysis	Provides guidance on threshold selection	May require domain-specific knowledge for threshold selection
[49]	Discusses the reconstruction of face mask images using auto-encoders	Challenges in face mask image reconstruction	Auto-encoders, face mask reconstruction	Addresses the reconstruction of face mask images	Limited information on the specific auto-encoder used
[50]	Highlights the use of encoder-decoder architecture in Deepfake generation	Challenges in Deepfake generation	Encoder-decoder architecture, Deepfake generation	Provides insights into the Deepfake generation process	May not cover all aspects of Deepfake generation
[51]	Introduces an innovative	Challenges in face	Adaptive GAN, face	Addresses limitations in	Limited information

	adaptive GAN for face transformations in Deepfakes	transformation using GANs	transformations	GAN-based face transformations	on the specific GAN architecture
[52]	Implements a CNN with multiple tasks for face recognition and orientation constancy	Challenges in face recognition and orientation constancy	Convolutional neural network, FaceNet	Aims to improve face recognition and orientation constancy	May require large labeled datasets for training
[53]	Discusses the use of Cycle GAN for constructing generative networks	Challenges in generative network construction	Cycle GAN, generative networks	Offers a method for constructing generative networks	Limited information on specific applications or datasets
[54]	Highlights concerns about the impact of Deepfakes on safety and privacy in democratic republics	Challenges in addressing the impact of Deepfakes	Deepfake impact, safety, privacy	Addresses a pressing societal concern	May not provide concrete solutions for addressing the impact
[55]	Recent advancements in DL enable automatic Deepfake detection	Challenges in automatic Deepfake detection	Deep learning, automatic detection	Addresses the need for automated detection	May require large computational resources for training
[58]	Implements Face swap-GAN to generate a Deepfake dataset	Challenges in generating a diverse Deepfake dataset	Face swap-GAN, Deepfake dataset	Provides a dataset for Deepfake detection research	May not cover all possible Deepfake variations
[59]	Produces low-budget, high-rating Deepfake movies using the VidTIMIT	Challenges in creating high-quality Deepfake movies	VidTIMIT dataset, Deepfake movies	Demonstrates Deepfake generation capabilities	Quality of Deepfake movies may vary

	dataset				
[60]	Deep learning techniques enhance image readability, face representation, and illumination	Challenges in improving image forensics with DL techniques	Deep learning, image forensics	Improves image forensics effectiveness	Limited information on specific DL techniques used
[66]	Focuses on detecting identity-swapped videos where one person's face is replaced with another's	Detection of identity-swapped videos	Auto-encoder architecture	Targets a specific type of deepfake manipulation	May not cover all types of deepfake manipulations
[67]	Investigates effectiveness of using specific facial regions for deepfake detection	Limited generalizability to other manipulation techniques	Deep learning, specific facial regions	Provides insights into effective feature regions	Limited to detecting identity-swapped videos
[68]	Integrates dynamic attributes from deep learning with hand-crafted features for facial expression detection	Effectiveness of hand-crafted features in deepfake detection	Deep features, hand-crafted features	Effective in identifying facial expressions	May not generalize well to all manipulation techniques
[69]	Utilizes frame-level data for deepfake detection using a multi-task learning network	Challenges in detecting deepfakes using frame-level data	Multi-task learning network, frame-level data	Effective in detecting deepfake manipulations	Limited interpretability of multi-task learning models
[70]	Focuses on detecting facial	Detecting facial manipulation	Long Short-Term Memory	Incorporates temporal information	May require substantial computationa

	manipulation using Long Short-Term Memory (LSTM) networks	in videos	(LSTM) networks	for detection	l resources
[71]	Uses 3D distance vectors between mouth landmarks, head rotations, and facial activity units for deepfake detection	Dependence on 3D landmarks for detection	3D distance vectors, mouth landmarks	Provides a unique approach using 3D information	Limited applicability to non-3D deepfake manipulation s
[72]	Combines deep learning with hand-crafted features to detect facial emotions and human behavior	Identifying facial emotions and behavior	Deep learning, hand-crafted features	Effective in identifying emotional and behavioral cues	May not generalize well to all manipulation techniques
[73]	Utilizes Multilayer Perceptron (MLP) for detecting deepfake videos based on eye blink patterns	Identifying deepfake videos based on eye blink patterns	Multilayer Perceptron (MLP), eye blink patterns	Achieves high accuracy in detecting specific manipulation s	May not cover all types of deepfake manipulation s
[74]	Investigates the use of peripheral facial features for detecting facial manipulation	Detecting facial manipulation using peripheral features	Peripheral facial features, 3D head position	Effective in distinguishin g manipulated features	Limited to specific types of manipulation s
[75]	Proposes a modified ResNet50 model for detecting	Detecting deepfake videos using modified ResNet50	Modified ResNet50 model	Achieves high accuracy in detecting deepfake	May require substantial computationa l resources

	deepfake videos			videos	
[76]	Investigates inconsistencies in eye blink patterns for detecting deepfake generators	Detecting deepfake generators based on eye blink inconsistencies	Long-Term Recurrent Convolutional Network (LRCN)	Achieves high accuracy in identifying inconsistencies	May not generalize well to non-eye-related manipulations
[77]	Uses an LSTM network to detect anomalies across frames for identifying movies with identity swaps	Detecting movies with identity swaps	LSTM network, frame-level features	Provides a unique approach using temporal information	Limited information on dataset diversity
[78]	Evaluates temporal action detection techniques in detecting deepfake videos	Detecting deepfake videos using temporal action detection	Temporal action detection techniques	Achieves superior performance in detecting deepfakes	- May require substantial computational resources
[79]	Investigates mesoscopic features in photographs using deep learning networks	Detecting mesoscopic features in photographs	Deep learning networks, mesoscopic features	Provides insights into mesoscopic feature detection	Limited interpretability of deep learning models
[80]	Uses a two-channel network to detect altered facial images	Detecting altered facial images using two-channel network	Two-channel network, facial alteration detection	Effective in detecting alterations in facial images	Limited information on model architecture
[81]	Develops a multi-task learning network for detecting deepfake videos	Detecting deepfake videos using multi-task learning	Multi-task learning network, deepfake detection	Improves performance in deepfake detection	May require substantial computational resources
[82]	Demonstrates the ability	Identifying deceptive	Capsule networks,	Outperforms other	Limited generalizabili

	of capsule networks to identify deceptive images and videos	images and videos	deceptive image, and video detection	algorithms in identifying deception	ty to non-deceptive manipulation
[83]	Investigates the use of specific facial regions as input to a pre-trained Xception network for deepfake detection	Detecting deepfakes using specific facial regions	Xception network, specific facial regions	Achieves high accuracy when using entire face input	Limited to specific types of manipulations
[84]	Uses an LSTM network to detect anomalies across frames for identifying movies with identity swaps	Detecting movies with identity swaps	LSTM network, frame-level features	Provides a unique approach using temporal information	Limited information on dataset diversity
[85]	Evaluates temporal action detection techniques in detecting deepfake videos	Detecting deepfake videos using temporal action detection	Temporal action detection techniques	Achieves superior performance in detecting deepfakes	May require substantial computational resources
[86]	Evaluates 3D CNN techniques in action detection for deepfake videos	Detecting actions in deepfake videos	3D CNN techniques, action detection	Achieves high accuracy in action detection	Limited information on model architecture
[87]	Proposes ID3, a deepfake detection model, and achieves high accuracy	Detecting deepfakes using ID3	ID3 model	Achieves high accuracy in deepfake detection	Limited interpretability of deep learning models

[88]	Proposes a methodology for distinguishing authentic movies from deepfakes	Distinguishing authentic movies from deepfakes	Methodology, deepfake detection	Provides a systematic approach to differentiate	May not cover all types of deepfake manipulations
[89]	Utilizes deep learning networks for feature analysis	Challenges in feature extraction and classification	Deep learning networks, feature analysis	Incorporates deep learning for improved accuracy	Complexity in feature extraction
[90]	Sets strong baseline for facial manipulation detection using modern DL architectures	Limitations of traditional handcrafted approaches	Deep learning architectures, modern DL techniques	Achieves robust detection using DL architectures	Lack of interpretability in DL models
[91]	Achieves impressive accuracy exceeding 97% in identifying manipulated video fragments	Addressing the proliferation of deepfake videos	Convolutional and recurrent neural networks (CNN and RNN)	Accurate detection of manipulated video fragments	Limited information on dataset diversity
[92]	Integrates temporal, behavioral biometrics based on facial emotions and head movements	Robust defense against deceptive media manipulations	Temporal and behavioral biometrics, facial emotions	Offers a multi-modal approach for detection	Limited robustness against adversarial attacks
[93]	Remarkable versatility in detecting various forms of attacks, accuracy rate exceeding 90%	Safeguarding against digital image and video manipulations	Capsule Networks, multi-modal approach	High accuracy in detecting diverse attacks	May lack scalability for large datasets
[94]	Importance of fine-	Identifying manipulation	Fine-tuning operations,	Provides insights into	Generalizability to all

	tuning operations, insights into identifying specific manipulation techniques	techniques in real-world scenarios	insights into techniques	real-world scenarios	manipulation techniques
[95]	Ensemble approach combining various CNN models, promising detection accuracy on extensive video datasets	The challenge of detecting manipulated faces in video sequences	Ensemble of CNN models, extensive video datasets	Achieves promising detection accuracy	May require significant computational resources
[96]	Combines geographical and temporal information, provides a robust defense against advanced manipulation techniques	Detecting imperceptible manipulations	Geographical and temporal information, robust defense	Excellent detection accuracy against advanced manipulations	May not cover all types of manipulations
[97]	Introduces the Set Convolutional Neural Network (SCNN) framework, achieving state-of-the-art results	Addressing the growing concern of misinformation in online content	Set Convolutional Neural Network (SCNN)	Achieves state-of-the-art results	May require substantial computational resources
[98]	State-of-the-art results in detecting Deepfake, Face2Face, and FaceSwap manipulations	Detecting misinformation in online video content	Recurrent convolutional models, various manipulation techniques	Effective detection of manipulated faces	May not generalize well to all manipulation techniques
[99]	Promising	Identifying	3D CNN	Robust	Complexity

	results with 3D CNN models exhibiting strong detection accuracy across various manipulation techniques	manipulated films across different techniques	models, various manipulation techniques	detection across diverse manipulation techniques	in training and deployment
[100]	Achieves adaptability even to previously unseen manipulation techniques	Identifying image and video modifications	Convolutional Neural Network (CNN)	Demonstrates adaptability to new manipulation techniques	Limited interpretability of CNN models
[101]	Trained forgery detectors achieve remarkable accuracy on the benchmark	Detecting various forms of facial manipulation	Trained forgery detectors, comprehensive benchmark	Provides a strong benchmark for evaluation	May not cover all emerging manipulation techniques
[102]	Achieves high accuracy, especially on publicly available datasets	Detecting fake videos using compact network architectures	Modified AlexNet model, publicly available datasets	High accuracy in detecting fake videos	Limited information on model architecture
[103]	Detects more than 98% of Deepfake videos, more than 95% of Face2Face videos	Challenges in detecting face tampering in videos	Compact network architectures, deep learning	High detection accuracy for specific manipulation techniques	May not generalize well to all manipulation techniques
[104]	Combines CNN with a fake face detector, achieving an excellent average accuracy	Identifying false face photos created using different modification techniques	CNN, fake face detector, various modification techniques	Excellent average accuracy in identifying manipulated photos	Complexity in model training and implementation
[105]	Promising	Detecting	Optical flow	Addresses	May require

	results in utilizing temporal dissimilarities for deepfake video identification	deepfake video based on temporal differences	fields, temporal dissimilarities	temporal dissimilarities for deepfake detection	computational resources for optical flow computation
[106]	Focuses on the study of multiple temporal parts together, uses Linear Support Vector Machines (SVMs)	Classifying data using SVMs for finding fake material	Spatial-temporal texture, Linear Support Vector Machines (SVMs)	Effective in finding differences between real and fake sequences	May not generalize well to all manipulation techniques
[107]	OC FakeDect achieves remarkable accuracy of 97.5% on the NeuralTextures dataset	Robust Deepfake detection using one-class anomaly approach	One-class anomaly detection, robust Deepfake detection	High accuracy in detecting Deepfakes	Limited to specific dataset and Deepfake variations
[108]	FReTAL excels in domain adaptation challenges, shows accuracy of up to 86.97%	Reducing the risk of catastrophic forgetting in deepfake detection	Transfer learning, domain adaptation, ReL, KD	Improves domain adaptation in deepfake detection	Limited interpretability in complex deep learning models
[109]	State-of-the-art results on three variants of the FF++ dataset, combines SCL with adaptive frequency feature generating module	Addressing the need for more discriminative and data-driven techniques	Frequency-aware feature learning, SCL, discriminative feature learning	Achieves state-of-the-art results on challenging datasets	May require substantial computational resources for training
[110]	Emphasizes the importance	Detecting subtle image manipulation	Spatial-Phase Shallow Learning	- Focuses on detecting subtle image	Limited information on SPSL

	of phase spectrum in detecting up-sampling artifacts	s using spatial-phase information	(SPSL), phase spectrum	manipulations	implementation details

PART 3

TYPES OF DIGITAL FACE MANIPULATIONS

3.1. ENTIRE FACE SYNTHESIS

One of the initial methodologies in this context was Progressively Generative Adversarial Network (ProGAN) [113]. The fundamental concept involved commencing with a lower resolution and subsequently incorporating more layers that imitate increasingly intricate characteristics during training, thereby gradually enhancing the synthesis process. Promising results were observed for synthesizing whole faces through experiments utilizing the CelebA dataset [114]. The source code for the ProGAN architecture is publicly available and may be accessed on the GitHub platform. Style Generative Adversarial Network (StyleGAN) was introduced as an improved iteration incorporating an alternate G architecture inspired by the style transfer literature [115]. The StyleGAN framework introduces a novel generator architecture that enables the manipulation of synthesis at different scales intuitively. This approach leads to the automatic acquisition of unsupervised separation between high-level attributes, such as pose and identity, in the case of human faces and the stochastic variation present in the generated images, such as freckles and hair. Illustrative instances of these modifications are depicted in Figure 3.1, explain generated during the training of the StyleGAN model using the CelebA-HQ and FFHQ datasets [116].

Figure 3.1. Examples of the entire face synthesis manipulation group, real images are extracted from http://www.whichfaceisreal.com/ and fake images from https://thispersondoesnotexist.com.

Lastly, it is worth mentioning two widely recognized GAN techniques, namely StyleGAN2 [117] and StyleGAN2-ADA [118], which incorporate adaptive discriminator augmentation. Insufficient data during the training of a GAN sometimes results in the phenomenon of discriminator overfitting, which subsequently leads to training divergence. The proposed approach of StyleGAN-adaptive discriminator augmentation introduces an adaptive discriminator augmentation mechanism, which effectively enhances training stability in scenarios with limited data availability. The proposed methodology allows for training a GAN without altering the loss functions or network topologies. This approach may be applied to either train a GAN from its initial state or to enhance the performance of an existing GAN when presented with a novel dataset. The researchers have provided evidence to support the notion that excellent outcomes can be attained even with a very small quantity of training photographs, specifically a few thousand. To conduct tests on detecting real and fraudulent images within this digital modification group, researchers must get real facial photographs from various publicly accessible digital repositories. These sources include CelebA [114], FFHQ [115], CASIA-WebFace [119], VGGFace [120], and MegaFace [121].

The Diverse Fake Face Dataset marks its recent entry into the field. To do full-face synthesis, the authors employed many synthetic pictures created by pre-trained models like ProGAN and StyleGAN [122].

The iFakeFaceDB dataset has been made available to the public [122]. The generation of the 250,000 and 80,000 synthetic facial images in this dataset was achieved using StyleGAN and ProGAN, respectively. To enhance its differentiation from prior dataset and maintain high realism while avoiding false detections, the iFakeFaceDB dataset employed a technique known as GANprintR (GAN fingerprint Removal) to eliminate the fingerprints of GAN designs. Figure 3.2 depicts two instances of fake photos, whereby one is generated directly by StyleGAN and the other is enhanced by eliminating the GAN fingerprint information. Including the GANprintR step in iFakeFaceDB renders it more formidable for robust fake identification than another dataset [117,118].

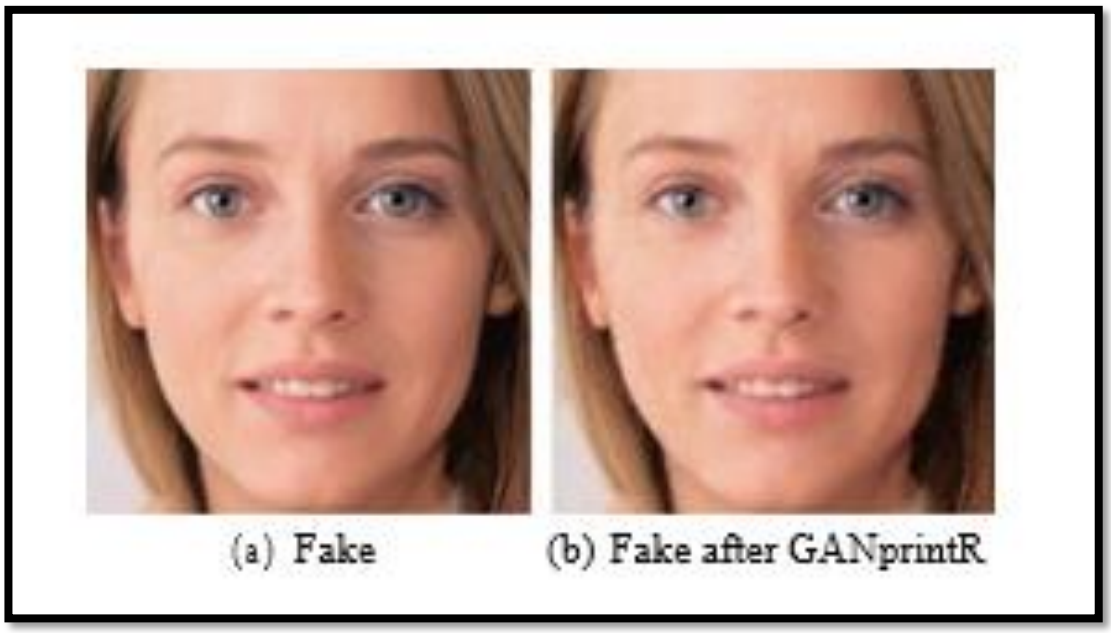

Figure 3.2. Examples of a fake image created using StyleGAN [122].

3.2. IDENTITY SWAP

To create the desired impact (target), the countenance of an individual depicted in the original footage is substituted with the visage of another individual. In contrast to the

thorough synthesis of facial features, the objective of identity swap is to provide fake videos that are highly persuasive. Figure 3.3 illustrates examples of visual data retrieval [123].

Figure 3.3. Examples of the Identity Swap manipulation group [123].

The video clips studied are authentic recordings extracted from YouTube, depicting individuals of different age groups, ethnicities, and genders, all belonging to the Celeb-DF dataset [101]. YouTube also offers a variety of high-quality and authentic representations of these transitions. This form of manipulation has the potential to yield benefits across all sectors, with the film industry being particularly poised to reap significant advantages. Conversely, this technology can be used for a wide range of illicit activities, including but not limited to financial fraud, identity theft, the creation of explicit footage of celebrities, and the dissemination of misinformation. Typically, two primary tactics are taken into consideration when engaging in the manipulation of identity swaps [124].

In computer graphics, conventional techniques such as FaceSwap coexists with innovative DL approaches like DeepFakes, accompanied by established software

packages such as FaceSwap and DeepFaceLab. The procedural stages associated with producing an identity exchange film are illustrated in Figure 3.3 [123].

The progression of visual changes in fake movies has been observed, beginning with the utilization of publically accessible fake dataset such as the UADFV dataset [125], and subsequently advancing to more recent dataset like Celeb-DF, Deepfake Detection Challenges (DFDC), DeeperForensics-1.0, and WildDeepfake datasets [126]. These modifications have contributed to the enhancement of realism in fake videos. Consequently, two discrete generations of identity swap dataset have emerged as shown in Figure 3.4.

The initial cohort comprises three distinct datasets. The UADFV dataset can be seen as an early instance of a dataset that was made available to users without any cost or restrictions. This compilation involved the digital superimposition of Nicolas Cage's facial features onto the original subjects in a total of 49 YouTube videos, utilizing the mobile application known as FakeApp. Consequently, the fabrication of all fake videos solely revolves around the portrayal of a single individual's identity. Each video segment has a duration of 11.14 seconds and showcases a solitary human with a resolution of 294 by 500 pixels [127].

An additional Deepfake dataset was generated by using the VidTIMIT database videos for the generation of Deepfake content [128,129]. The VidTIMIT collection includes 620 created videos that cover 32 subjects [130]. The creation of fake videos has been made more accessible by using a commonly used GAN face-swapping technique.

The GAN face-swapping technique makes use of a generative network produced by CycleGAN [131] and the weight parameters of FaceNet [132]. To achieve accurate face matching and reliable face recognition, researchers have created a Multi-Task Cascaded Convolutional Neural Network [133]

Figure 3.4. Graphical representation of the weaknesses presents in Identity Swap [125]; Celeb-DF and DFDC (2nd generation) [126].

3.3. FACE MORPHING

To generate fake biometric face samples that exhibit characteristics like the biometric information of many individuals, one can employ digital face alteration techniques such as face morphing [134]. The development of morphing face pictures raises concerns about the reliability of facial sample checks used in various security systems. Studies have shown that these pictures have the potential to bypass such

checks for multiple individuals [135], posing a significant threat to the effectiveness and accuracy of face recognition systems. The illustration presented in Figure 3.5 exemplifies the utilization of digital editing techniques to morph a facial image [134]. It is important to note that face morphing primarily focuses on image-level modifications, such as identity swaps, rather than video-level transformations. As depicted in Figure 3.5, it is common to consider the frontal view of the face as well. Extensive research has been conducted about face morphing in recent times.

In previous studies, extensive investigations have been conducted on both morphing strategies and morphing attack detectors [135]. The generation method of face morphing images often involves considering the following three phases in sequential order. Identifying shared characteristics among the multitude of shown visages. Several commonly used techniques for achieving this objective involve:

- The isolation of identifiable facial attributes such as eyes, noses, mouths, and so on.
- The manipulation of the original facial images to ensure that their corresponding elements (known as landmarks) are geometrically aligned within the samples.
- The merging of color values from the altered images. Post-processing techniques are frequently employed to eliminate anomalous artifacts resulting from pixel or region-based morphing [136].

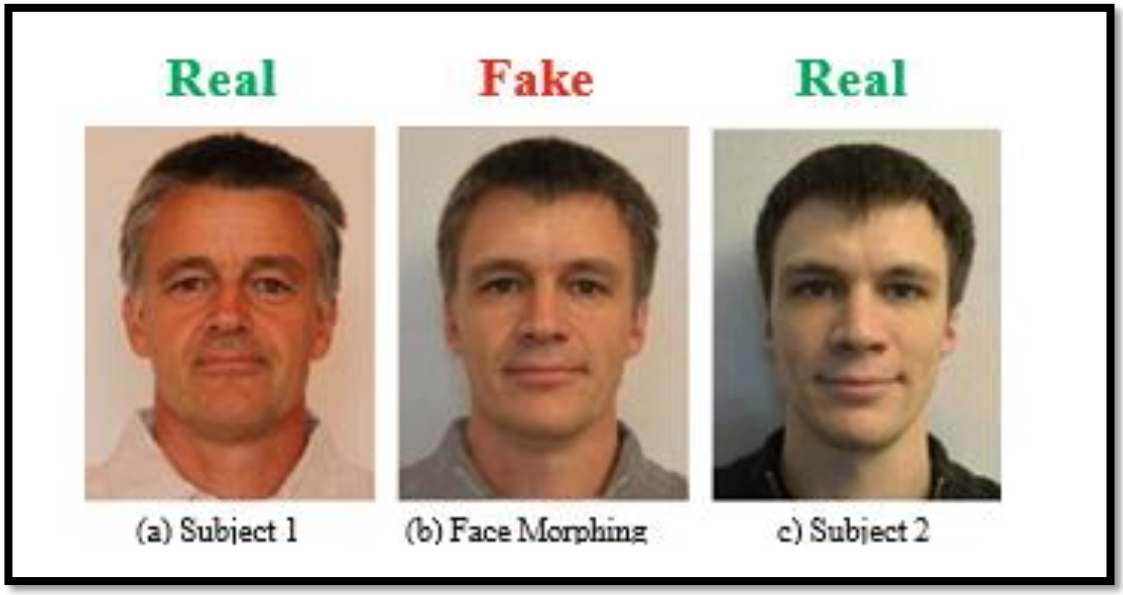

Figure 3.5. Example of a face morphing [134].

3.4. ATTRIBUTE MANIPULATION

Facial variations encompass a range of modifications, including changes in colour, hair pigmentation, chronological age, biological sex, and the utilization of eyeglasses [137]. The technique of modification is frequently implemented by employing GANs, such as the previously introduced StarGAN approach [138]. An example of such manipulation is demonstrated by the extensively utilized smartphone application recognized as FaceApp. The emergence of this technology has rendered it possible to practically replicate the act of trying on eyeglasses, applying makeup, and even exploring various haircuts. The instances of attribute adjustments created by FaceApp are depicted in Figure 3.6, as documented by reference [139].

The application of GAN frameworks for modifying facial characteristics is prevalent [140]. Nevertheless, the researchers faced a dearth of publicly accessible information pertaining to this specific field, which posed a challenge for their research endeavors. One of the primary factors that contributes significantly to this phenomenon is the extensive accessibility of source code for various GAN techniques on the Internet. The provided accessibility facilitates the seamless generation of customized synthetic datasets by researchers to cater to their individual requirements. This section provides a compilation of the most recent approaches in the field of Generative

Adversarial Networks (GANs), organized in reverse chronological order. Additionally, matching references to source code are included.

Figure 3.6. Examples of the Attribute Manipulation group [137].

The integration of an encoder with a conditional GAN to develop the Invertible Conditional GAN enables sophisticated picture manipulation at a higher level. This methodology yields precise results when altering properties. However, this alteration significantly transforms the visual aspect of the object [141].

An encoder-decoder framework can be effectively trained to reconstruct images by capturing the essential features of the image and the attribute values directly in the latent space. Nevertheless, like the Invertible Conditional GAN, the generated images may have significant omissions of essential elements or exhibit unforeseen distortions [142].

3.5. EXPRESSION SWAP

Face reenactment refers to the process of modifying the facial expression of an individual. While the literature proposes various techniques for manipulation, such as utilizing well-known GAN architectures for image-level manipulation [84], our research group specifically concentrates on two widely recognized methods, namely

Face2Face and NeuralTextures [143]. These methods involve substituting the facial expression of one individual in a video with that of another. The screenshots from the FaceForensics++ dataset [144] are depicted in Figure 3.7. The utilization of such manipulative tactics may yield significant ramifications, as evidenced by the dissemination of a viral film wherein Mark Zuckerberg is observed providing misleading statements.

The only publicly available dataset for research in the field is an extended version of FaceForensics known as FaceForensics++ [145]. The primary focus of the FaceForensics dataset was first directed towards the Face2Face approach. The proposed technique in computer graphics enables the transfer of emotional content from one video to another while preserving the subject's identity [146].

In every video, the initial frames were used to create a transient three-dimensional model of the individual's face, afterward employed to monitor the individual's facial expressions throughout the remainder of the film. Subsequently, the set of 76 Blendshape coefficients representing the source expression parameters for each frame was used in the application of the target video, resulting in the production of simulated iterations of the initial content. In a subsequent iteration of FaceForensics++ [143], a novel learning methodology centered around NeuralTextures. The approach employed in this study involves utilizing the raw video data to train a neural network that captures the texture of the subject. This neural network is integrated into a rendering network. The authors used the patch-based GAN-loss of Pix2Pix in their approach [146].

Figure 3.7. Examples of the expression swap manipulation group [144].

The chapter explores various types of digital face manipulations, including entire face synthesis, identity swap, face morphing, attribute manipulation, and expression swap. These manipulations encompass a wide range of techniques and applications, from generating high-resolution images using Progressive Generative Adversarial Networks (ProGAN) to altering facial attributes like complexion and hair pigmentation using Generative Adversarial Networks (GANs) such as StarGAN and FaceApp. The discussed techniques have both positive and negative implications. On the positive side, technologies like face morphing and attribute manipulation can be used for creative purposes in the fashion and entertainment industries, allowing users to experiment with different looks and styles. Additionally, identity swap and expression swap techniques have potential applications in filmmaking and special effects. However, these technologies also pose significant challenges and risks. The ability to create highly convincing fake videos and images, such as those generated by identity swap and expression swap methods, raises concerns about misinformation, identity theft, and financial fraud. Face morphing can undermine the reliability of facial recognition systems, potentially compromising security measures.

The availability of publicly accessible datasets, such as CelebA and FFHQ, has facilitated research in this area. Researchers have developed datasets like Celeb-DF and FaceForensics++ to support the study of deepfake detection and manipulation detection techniques. These datasets play a crucial role in evaluating the effectiveness of detection methods.

In conclusion, digital face manipulations have become increasingly sophisticated, thanks to advancements in deep learning and GAN technology. While these techniques offer exciting possibilities for creative expression and entertainment, they also present significant challenges and risks in terms of privacy, security, and the spread of misinformation. The development of datasets like Celeb-DF and FaceForensics++ has allowed researchers to explore and develop detection methods to identify manipulated content accurately. These efforts are crucial in mitigating the potential harm caused by the misuse of digital face manipulation technologies. As technology continues to evolve, it is essential to strike a balance between innovation and responsible use. Ethical considerations, privacy safeguards, and detection mechanisms must be continually developed to address the challenges posed by digital face manipulations. Additionally, public awareness and education regarding the existence and implications of deepfake technology are essential to empower individuals to discern and respond to manipulated content effectively.

PART 4

METHODOLOGY

This chapter explains the methodology applied throughout this research. Section 4.1 explains our RACNN model and datasets. Section 4.2 explains a comparative study of three enhancement-based models. Then, Section 4.3 explains the performance metrics. Finally, Section 4.4 explains transfer learning techniques for neural networks.

4.1. RATIONALE-AUGMENTED CONVOLUTIONAL NEURAL NETWORKS (RACNN) MODEL

Using the RACNN models applied in this study, our study found out that the Deepfake facial reconstruction problem has already been solved in terms of accuracy and for most available datasets. Figure 4.1 is a flowchart illustrating this process. Alternatively, our models were robust enough to perform Deepfake facial reconstruction without a publicly available training/testing dataset.

.

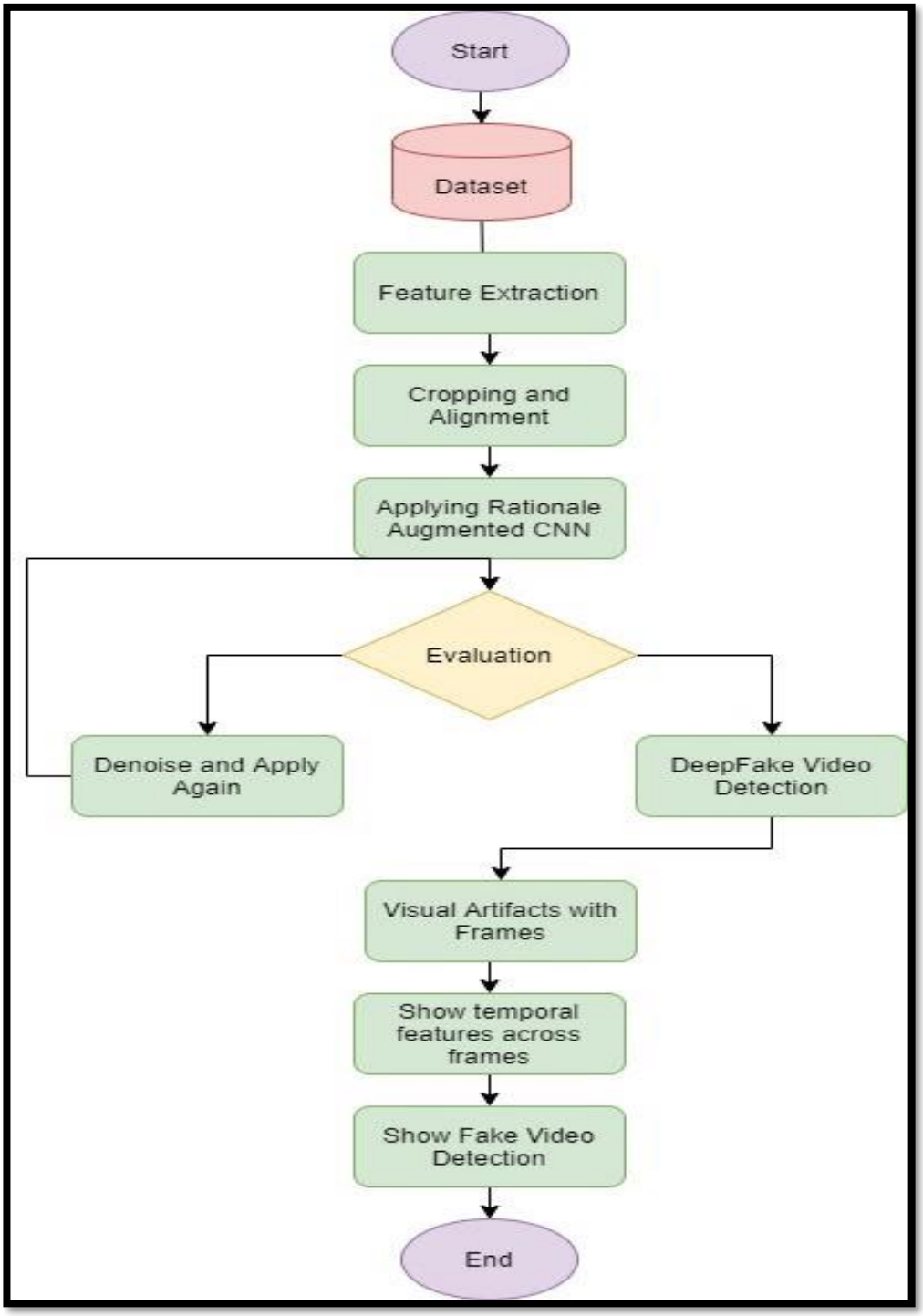

Figure 4.1. Flowchart of the procedural method.

4.1.1. Dataset Description

This section explains the two data sets used in our RACNN model.

4.1.1.1. Deepfake Detection Challenges Dataset

The Deepfake Detection Challenges dataset is a critical resource for addressing the emerging problem of deepfake videos, which are manipulated videos created using advanced AI techniques. This dataset has been carefully curated to include a wide range of real and Deepfake videos of varying complexity. It typically includes samples from a variety of sources, including different facial expressions, lighting conditions, camera angles, and scenarios [145].

One of the primary goals of this dataset is to facilitate the development and evaluation of deepfake detection models. Researchers and practitioners use this dataset to build and refine machine learning algorithms that can effectively distinguish between authentic and manipulated videos. The dataset is organized into training, validation, and test subsets, allowing models to be trained on a large set of labeled data while evaluating their performance on unseen examples. This approach supports the development of robust and accurate deepfake detection algorithms that can potentially mitigate the harmful effects of misinformation and fraudulent content spread through manipulated videos.

4.1.1.2. Faceforensics++ Dataset

The Faceforensics++ dataset, or FF++, dataset is a significant contribution to the field of Deepfake detection research. This dataset is specifically designed to advance the capabilities of Deepfake detection methods. It consists of an extensive collection of manipulated videos that simulate real-world scenarios in which facial manipulation occurs. The dataset includes various manipulation techniques such as face swapping, reenactment, and more to generate convincing deepfake content [146].

Faceforensics++ provides a comprehensive suite of evaluation videos spanning multiple domains, including news, entertainment, and social media. It also provides corresponding original and manipulated videos, providing a controlled environment for training and testing deepfake detection models. Researchers and data scientists

can use this dataset to develop sophisticated deepfake detection methods using CNNs, attention mechanisms, and other state-of-the-art techniques. The diversity and realism of the dataset helps create robust and accurate models that can detect even the subtlest signs of facial manipulation, promoting trust in digital media and protecting against the potential misuse of manipulated content.

In our RACNN model, these datasets collectively serve as the basis for training, 99260 images belonging to 2 classes with 70%, 1030 images belonging to 2 classes with 10% validating, 26914 images belonging to 2 classes with 20% for testing and fine-tuning your Deepfake detection algorithms, ensuring that your model can effectively distinguish between real and fake videos, ultimately contributing to the ongoing effort to combat the spread of deceptive digital content.

4.1.2. Feature Extraction

Feature extraction using a RACNN involves exploiting the network's ability to highlight important regions, or "rationales," within an input image or video frame. These highlighted regions can serve as meaningful features for downstream tasks such as classification, object detection, or further analysis. Here's how feature extraction with a RACNN typically works:

1. Input Data: Start with the input video frame.

2. Rationale Generation: Pass the input data through the RACNN. The RACNN employs advanced techniques such as attention mechanisms to identify and highlight specific regions of interest within the input.

3. Feature Extraction: Extract the highlighted regions or rationales as features. These regions are often represented as spatial maps that indicate the importance of each pixel or area within the input.

The key advantage of using a RACNN for feature extraction is that it provides meaningful and context-aware regions of interest, which can enhance the

performance of subsequent tasks. This is especially valuable in scenarios where transparency and interpretability are important, as the rationale behind the features is readily available.

4.1.3. Cropping and Alignment

Cropping and alignment are essential preprocessing steps in deepfake detection, as they help ensure that input images or frames are in a standardized format, making it easier to compare and analyze them. Here's how cropping and alignment are typically performed in deepfake detection:

1. Face Detection and Localization:
   - Initially, face detection algorithms are used to locate faces within an image or video frame.
   - These algorithms identify facial landmarks and the bounding box around the detected face(s).

2. Cropping:
   - Once the faces are detected, a cropping process is applied to isolate the face region. This involves extracting the pixels within the bounding box.

3. Alignment:
   - To ensure consistent facial features across different frames or images, alignment techniques are employed.
   - Facial alignment involves adjusting the position and orientation of the detected face within the bounding box.
   - Common alignment methods include landmark-based alignment, where facial landmarks (e.g., eyes, nose, mouth) are used to normalize the face's position, rotation, and scale.

4. Normalization:

- After cropping and alignment, the facial region is often normalized to a fixed size and orientation. This step is crucial for ensuring that all faces are represented consistently for analysis.

- Normalization may involve resizing the cropped face to a predefined resolution (e.g., 224x224 pixels) and ensuring it's centered within the image.

Figure 4.2 explains that by cropping and aligning the faces within input images or frames, Deepfake detection models can focus on the most critical information, which is the facial region. This preprocessing step improves the model's ability to detect inconsistencies and anomalies that may indicate the presence of a Deepfake.

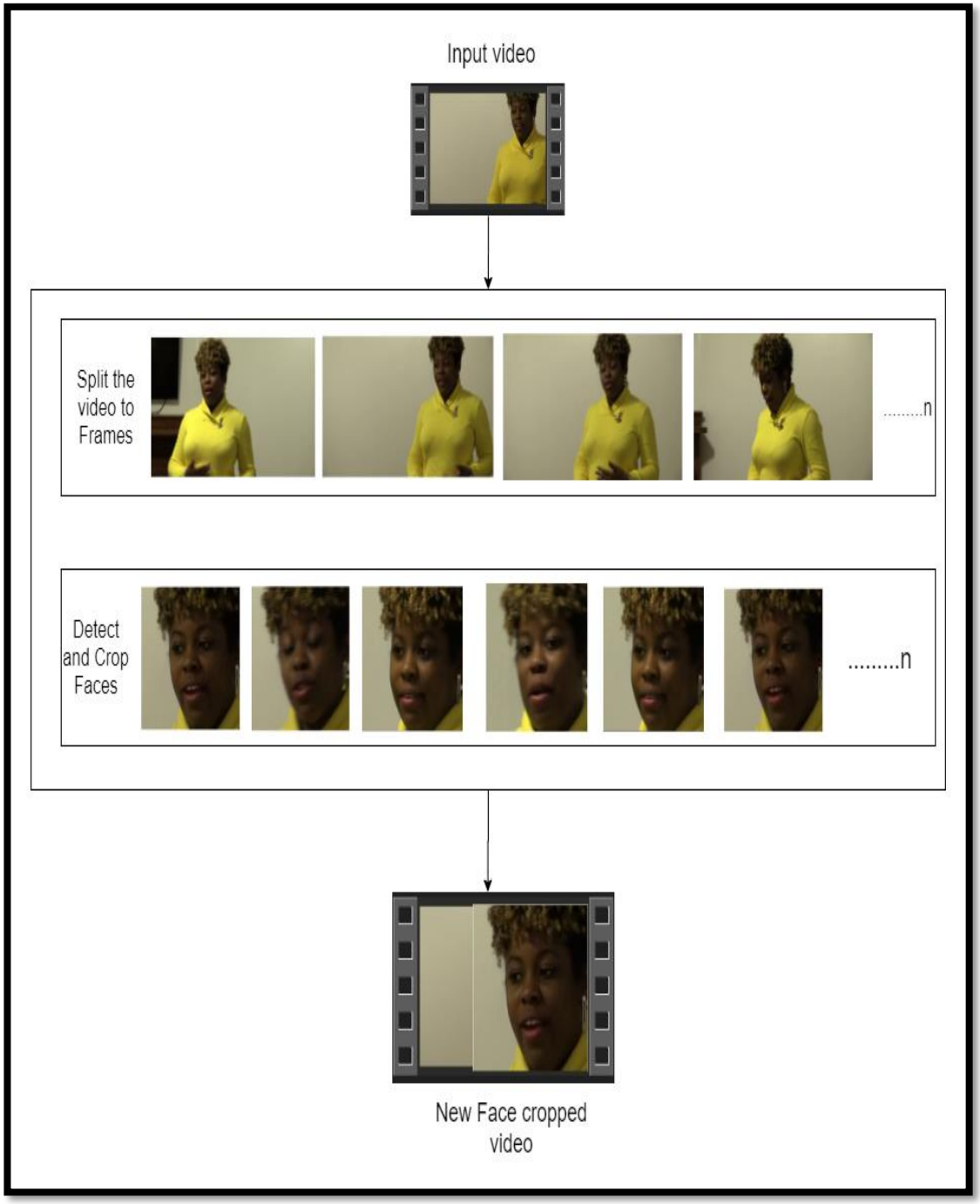

Figure 4.2. Cropping and Image alignment.

4.1.4. Donald Trump Filter with RACNN Model

In our study, we utilize the "Donald Trump Filter," a specific computational filter characterized by its distinct attributes and capabilities that have been instrumental in facilitating the detection process in the context of our thesis. Rationale augmentation in CNNs is a technique used to improve the interpretability and explain ability of the network's decisions.

The RACNN content was structured based on both the VGG model and the ethnic background of the individuals. The binary classification study resulted in identifying two categories: Category 0 includes real facial images, which include natural, legitimate, and disguised samples, while Category 1 refers to forgeries, which includes impersonator face images.

Using the DFDC and FF++ datasets [123], all three models were trained for ten iterations at a training rate of 0.001 over ten hours. We applied the dataset to the test set to determine its reliability. We used data augmentation to flip all the original images horizontally and vertically, thereby doubling the growth rate (original image plus horizontally flipped image plus vertically flipped image).

The Custom CNN design connected six convolutions via batch normalization, maximum pooling, and dropout layers (Conv2D). The input and output layers use a sigmoid function. However, the input layer employs a Rectified Linear Unit (ReLU). To improve the accuracy of the image analysis, we added padding to one kernel, and the method was applied to all layers to prevent overfitting. We trained and validated custom architectures on the original and enhanced datasets at a scale of 1/255. We created a training set to understand how data aggregation affects simulation results. Since not all photos had the same pixel quality, we included a horizontal flip, a 0.2 zoom range, a 0.2 shear range, and a rescaling factor in the operation level to compensate for changes in image quality that would otherwise have affected the model's classification performance.

We modeled a 16-layer CNN architecture, consisting of pooling layers, five max-pooling layers, and one SoftMax layer using an approach identical to that of the VGG-19 model. Since VGG-19 has been pre-trained on numerous object classes, it

can acquire representations of deep features. VGG-19 has demonstrated the ability to correctly classify full faces. We applied a high-end setup to it by adding a dense layer after a previous layer block providing the face features and a dense network as the output layer to sigmoidal activation to create a model for Deepfake identification tasks.

This study extends the Keras DenseNet-264 architecture by introducing a denser output layer. This architecture involves the placement of a 2x2 convolutional layer after a 3x3 MaxPooling layer with a stride of two. Furthermore, the input layers are equipped with a sigmoid activation function, batch normalization, and the ReLU activation method. In addition, the ensemble contains four full cubes, and the transition layers connecting each dense block contain a two-by-two average max pooling along with a one-by-one convolution. The classification model developed for this dense block-centered output layer accumulates image features from all levels of the network and is positioned before the final dense block. This model's training set comprises one hundred thousand photographs, validated against an additional twenty thousand. The training and validation utilize scaled versions of the baseline, grayscale, and enhanced datasets, incorporating horizontal flips, 20-degree rotations, and the rescaling method applied in the customized CNN architecture to enhance training data diversity. A separate training of the DenseNet structure on grayscale-only data investigates the influence of color on classifying data into fake and real categories due to variations in pixel-level resolution between grayscale and color images.

However, the VGG framework is constructed and evaluated exclusively on the primary dataset. All models are built as expected, except for an additional deep network with a sigmoid activation function. This additional layer consistently provides a useful rectifier nonlinear activation for binary classification, providing probability outputs between 0 and 1 that can be easily translated into specific class values. Currently, implementing a facial reconstruction system for security requires training all new individuals, which is not currently feasible with the existing architecture. The computational cost of a single forward pass image precludes real-time face reconstruction. The training of the cross-entropy loss function, as shown in

Figure 4.3, guides the model to optimize the correct number of classifications within a dataset, rather than quantifying new facial features into an encoded vector. These two key challenges must be overcome to build a robust statistical model with high reliability and performance.

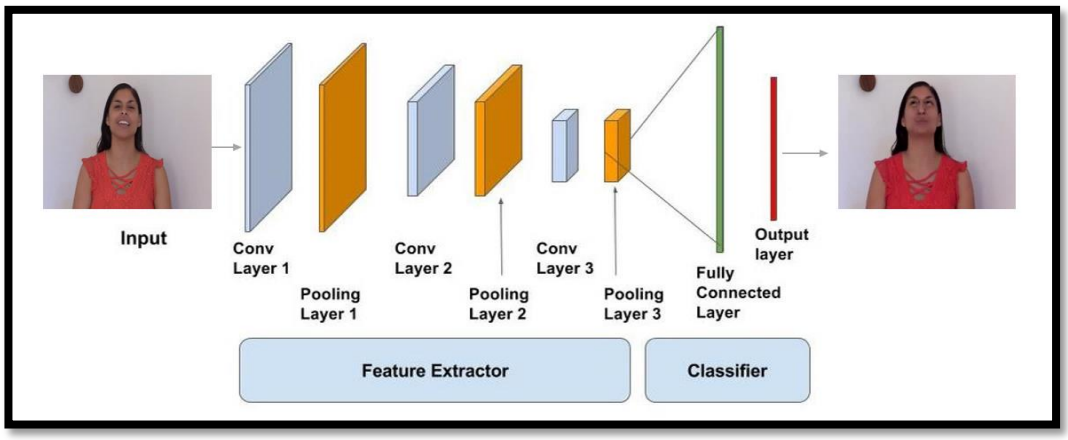

Figure 4.3. Schematic of a rationally enhanced CNN network.

It involves enhancing the model's capability to provide explanations or justifications for its predictions by highlighting relevant input features. Here are some key features of rationale augmentation in CNNs:

1. Interpretability: Rationale augmentation helps make CNNs more interpretable by generating explanations for why a particular prediction was made. This can be crucial in applications where understanding the model's reasoning is important.

2. Attention Mechanisms: Rationale augmentation often incorporates attention mechanisms into CNNs. These mechanisms allow the model to focus on specific regions or features of the input data that are most relevant to the prediction, making the model's decision more transparent.

3. Highlighting Important Features: Rationale augmentation methods aim to highlight important input features or regions that contributed most to the model's decision. This can be visualized as heatmaps or saliency maps, showing which parts of the input were influential.

4. Improved Trustworthiness: By providing rationales or explanations, CNNs with rationale augmentation can instill more trust in their predictions, especially in critical domains such as healthcare, finance, or autonomous systems.

5. Human-AI Collaboration: These models facilitate human-AI collaboration by enabling humans to understand and potentially correct model decisions when necessary. Users can intervene based on the provided rationales.

6. Reducing Bias and Errors: Rationale augmentation can help identify biases or errors in model predictions by making it easier to detect when the model is relying on irrelevant or biased information.

7. Natural Language Generation: In some cases, rationale augmentation may involve generating natural language explanations alongside the visual or numerical justifications, making the model's output more accessible to users.

8. Explainable AI: Rationale augmentation aligns with the broader field of Explainable AI, which aims to make AI models more transparent and understandable to humans.

Overall, rationale augmentation in CNNs enhances their transparency and the ability to provide meaningful justifications for their predictions, making them more useful and trustworthy in various applications.

Then temporal features in the context of video analysis refer to patterns and information that evolve over time across multiple frames of a video sequence. Analyzing these features can be crucial for tasks such as motion detection, object tracking, action recognition, and more. These methods for visualizing temporal features can be valuable in various video analysis applications, from surveillance and security to entertainment and sports analytics. Depending on the specific task, the choice of visualization technique and feature extraction method may vary. Finally show fake video detection.

4.2. A COMPARATIVE STUDY OF ENHANCEMENT-BASED MODELS

This section examines neural network techniques based on transfer learning. Transfer learning [149] involves employing pre-trained models in the prediction process, leveraging previously acquired knowledge to enhance prediction performance. The fine-tuning methods based on transfer learning retrain specific sections of pre-existing networks using new datasets.

This section analyzes the working principle of transfer learning approaches used for deepfake detection. The configuration settings and architecture of (NN) techniques are analyzed. This is followed by a classification of techniques with descriptions of each.

- **Proposed Net:** A CNN can be constructed by overlaying multiple neural subunits. In each epoch, weight values are updated in the training phase using techniques such as backpropagation, which establishes connections between neurons and assigns weights to their connections. The initial part of the CNN captures features, while the subsequent part handles categorization. Pre-trained networks such as DenseNet, accessible through the Keras API, have been employed. The DenseNet architecture, shown in Figure 4.5, includes models such as DenseNet201, DenseNet169, and DenseNet121, chosen to improve computational efficiency. The FF-CNN facilitates the interconnection between each level, ensuring that inputs from previous levels to maintain the feedforward nature, raising them for subsequent levels [150].
- **The Dense Block:** The Dense Block is a CNN module that directly connects all layers (with equivalent feature map sizes). Initially proposed as a part of the DenseNet design, it ensures feed-forward nature by receiving additional inputs from previous layers and distributing its feature maps to subsequent layers. Unlike ResNets, features are concatenated rather than summed before being added to a layer. Consequently, a layer with l inputs gathers feature maps from all prior convolution blocks and distributes them to subsequent Level L layers. This approach adds L(L+1)/2 connections to an L-layer network, referred to as "dense connectivity". Convolutional layers play a

critical role in neural networks by capturing complex features from input data of fixed dimensions. DenseNet architectures include multiple dense blocks, such as the DenseNet169 with 169 layers distributed across 4 dense blocks, each consisting of three transition layers, a classification layer, and a convolutional layer. This model is initiated with max pooling of 56 following a 112 convolution, accepting 224x224 RGB images as input. The architecture, shown in Figure 4.4, illustrates a dense block (DB) consisting of six stacked layers [149].

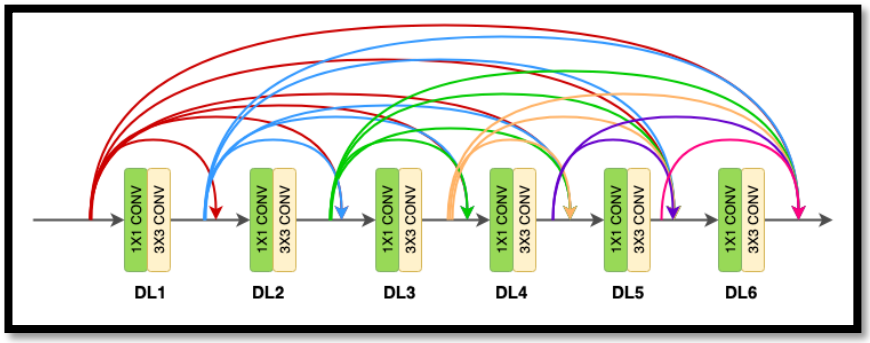

Figure 4.4. Dense block (DB) with dense layers (DL) [149].

- **Dense Connections:** Dense connections are a type of layer found inside a (DNN). This layer is also known as Fully Linked Connections because it is the only layer in which each input is weighted and connected to each output. This means that both input and output parameters are available. This can result in many parameters for a large network, as shown in Figure 4.5, which represents the dense connection process. Eq. (4.5) describes the calculation of dense connections [149].

$$h_l = g(W^T h_{l-1}) \tag{4.5}$$

where g is a function of activation.

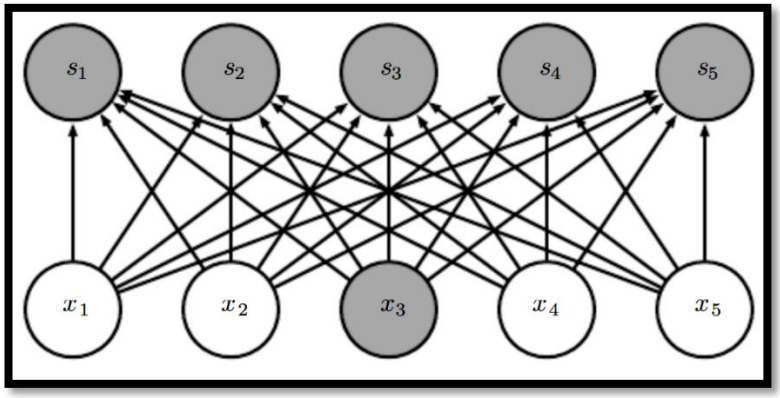

Figure 4.5. Dense Connections processes [149].

- **The DenseNet**: With dense blocks, we directly connect all layers (with equivalent feature map sizes). To maintain the feed-forward nature, each layer receives additional inputs from all previous layers and sends its feature maps to all subsequent layers. Figure 4.6 shows a dense net design [149].

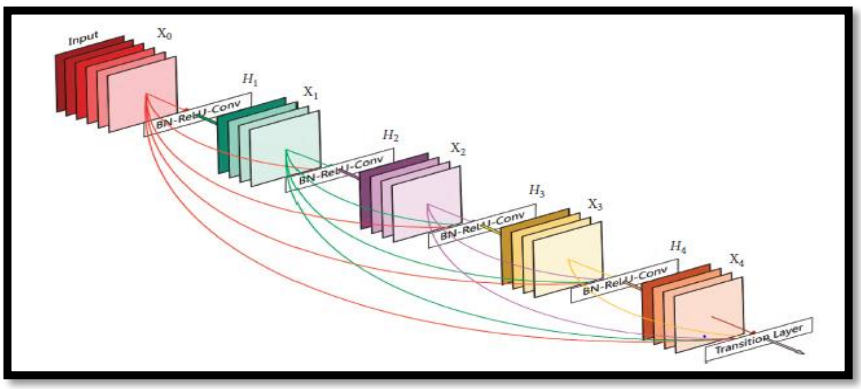

Figure 4.6. The shape of Dense-Net [127].

- **DenseNet121:** The architecture of DenseNet represents a significant advancement over Residual CNN (ResNet). Unlike ResNet and other CNNs, DenseNet has direct connections between each layer and the next within the network [127]. Notably, DenseNet121, a precise model within the Keras framework, demonstrates extensive use of dense layers similar to the final layers. The model's compact structure encompasses multiple tightly packed blocks, including levels such as Batch Standardization (BN) and 3 × 3 turnaround. Between the dense blocks, there are transition layers that contain

an average pooling of 2x2 and a 1x1 convolution. Following the previous dense block, a customized dense layer with sigmoid activation is incorporated. Furthermore, DenseNet-121 integrates six densely stacked layers within a dense block. The output of each dense layer correlates with its growth rate.

- **DenseNet201:** The DenseNet201 uses a compressed network to facilitate training simplicity and parameter-efficient models that are expected to recycle features across successive layers. As a result, the input to the next layer is more diverse, which improves performance [149].

- **DenseNet169:** One of the most effective models for dealing with fading gradients has 169 levels and only three parameters. In addition, ResNet-50 was integrated into the evaluation framework of this study for performance evaluation. The Residual Network (ResNet) is a neural architecture characterized by the incorporation of multiple deep layers to improve accuracy and efficiency in solving complex tasks. The introduction of more layers is based on the premise that this augmentation will increase the complexity of the layer attributes [149].

- **ResNet50:** ResNet50 consists of forty-eight convolution layers. This includes a maximum pool layer, and a typical pool layer. The sum of each of its floating-point processes is 3,8*109 [150]. This is shown in Figure 4.7. Instead of learning unsourced functions, ResNets learn residual features about the layer inputs. Instead of assuming that each of the few stacked layers directly corresponds to a preset underlying mapping, residual nets allow these stacked layers to adhere to a residual mapping. Residual nets are constructed by stacking residual blocks on top of each other. For example, a ResNet50 consists of fifty layers of these blocks. Formally, with the required base mapping represented as H(x), we let the stacked nonlinear layers accommodate an additional mapping of F(x): = H(x)x. The initial mapping is modified to become F(x+x). There is empirical evidence that such networks are easier to tune or can be optimized from a greater depth [150].

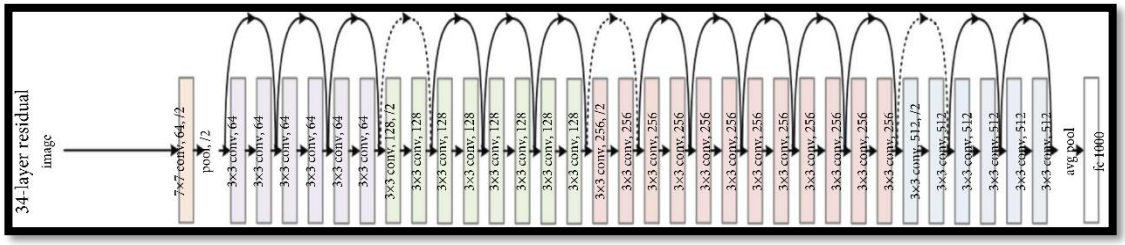

Figure 4.7. Design of Res-Net50 [150].

- **VGG16:** The widely used VGG16 model is primarily used for image recognition tasks. It uses a CNN structure and was first used in the 2014 Large Scale Visual Recognition Challenge. In this study, VGG16 was adapted for Deepfake detection by replacing the top layers with a multilayer perceptron (MLP) [129]. The architecture of the model is detailed in Table 4.1, where the first layer is the input layer. Subsequent layers include various components, including flattened layers, drop layers for regularization, dense layers for complex feature extraction, and sigmoid activation output layers for Deepfake classification. The focus is on 3*3 filter convolution levels and single pooling stages. The model culminates with a SoftMax output after two fully connected layers, where the "16" in VGG16 denotes the sixteen weighted layers in the model, as shown in Figure 4.8. Table 4.1 examines the configuration parameters and architectural layer analysis of the VGG16 model. The initial layer acts as an input layer with dimensions (100, 256, 256, 3). The implementation of VGG16 modeling layers with 512 units and 14,714,688 training parameters follows. A drop layer is a subsequent layer of architecture that prevents model overfitting. Using flattened layers, the pixel data is turned into a series of a one-dimensional arrays. Predictions of Deepfake depend on a series of complex layers. In the dense layer, there are six units with feedback connections. To categorize Deepfakes, we applied the output data layer with sigmoid activation. The emphasis was placed on 3*3 filter convolution levels and a single tempo rather than many hyperparameters. These stages used the same padding and maximum pool level as the 2*2 filter speed. Convolution and maximum pool levels are organized uniformly across the design. The output is a SoftMax after two

fully connected levels. The number 16 in VGG16 alludes to the fact that it contains sixteen weighted levels, as seen in Figure 4.8.

Table 4.1. Model configuration settings and layer structure analysis for VGG16.

#	Layers	Unity	Function of Activation	Shape of Output	Parameters
1	Sec. Layers of Input	/	/	(100, 256, 256, 3)	zero
2	Layers of VGG16	512	/	(None, 8, 8, 512)	14,714,688
3	Layers of Dropout	0.2	/	(100, 8, 8, 512)	zero
4	Layers of Flatten	/	None	(100, 32,768)	zero
5	Layers of Dense	64	Re-LU	(100, 64)	2,097,216
6	Layers of Output	1	Sigmoid (curved)	(100, 1)	65

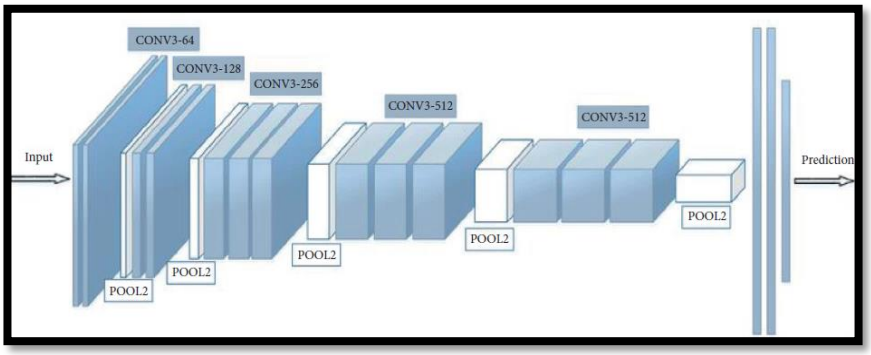

Figure 4.8. Design of VGG16 [151].

- **VGG19:** In terms of performance, a single convolutional layer outperforms the sophisticated CNN approach. For example, the layer allows max-pooling downsampling and a modified ReLU activation function to select the highest values within an image region for area averaging. Downsampling layers are often used to reduce variables while preserving important sample features, as shown in Figure 4.9 for VGG19.

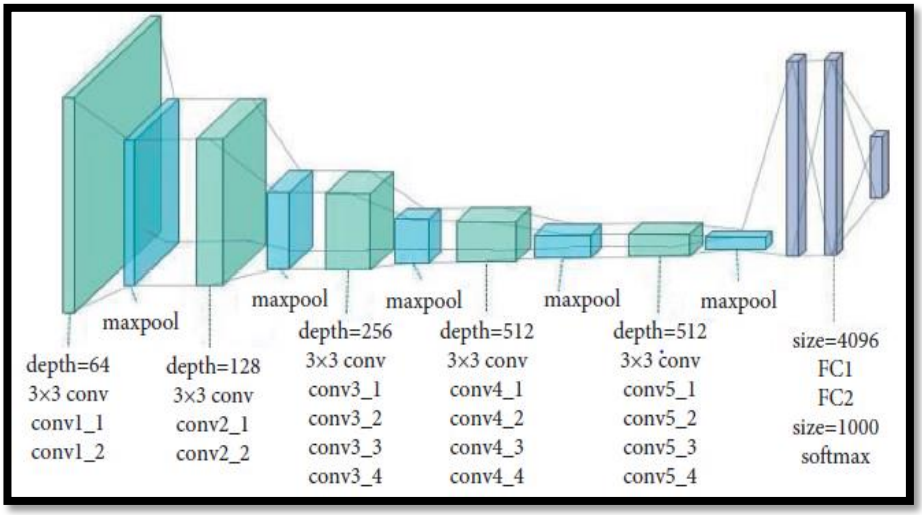

Figure 4.9. Design of VGG19 [151].

- **The VGG-Face:** The most efficient technique for mental image identification is to use the standard face recognition datasets published by the Oxford Visual Geometry Group [152]. This strategy allows us to create a large training dataset while consuming little energy. All five rows of the layer contain convolutional and max-pooling levels. Each first and second block had two 3x3 convolutional levels and one pooling level.

For the 3x3 convolution, the levels are determined by the max-pooling level, with each level consisting of blocks three, four, and five. At each convolution level, we used the ReLU activation function. VGG-Face requires pre-trained weights. Finally, we added dense levels to our five-level blocks to achieve the necessary facial features. The final step was to add the net production and the activated sigmoid to the dense level set. As shown in Figure 4.10, this study also introduced a pattern that shows the same overall difference.

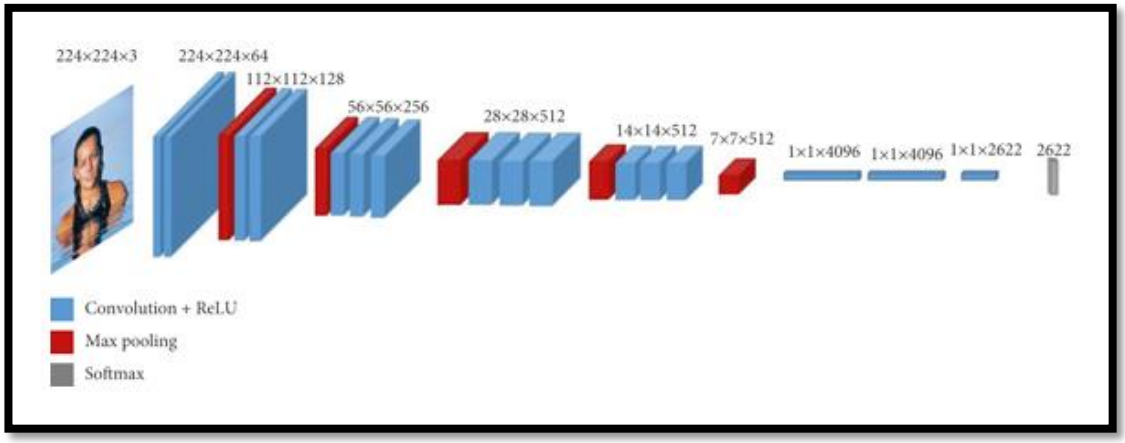

Figure 4.10. Design of VGG-Face [152].

They misclassified 138 real images as fake and 497 as false, making it hard to distinguish real images from phony ones. Figure 4.11 shows a screenshot of the classification of "true" and "false" images.

Figure 4.1. Classification of the real and fake pictures.

Once the face is detected and an image of a face is cropped, reconstruction with a triplet-loss trained network becomes a straightforward operation. A forward pass in such a model produces a 128-dimensional vector representing the unique identification of the individual. This result indicates the degree of similarity between two faces by calculating the L2 (Euclidean) distance between the 128-dimensional identity vectors of two photographs. In addition, a threshold value determines how similar or dissimilar a face must be to be considered a match. Comparing the 128-dimensional feature vector to a set of known encodings until another vector is found within the threshold makes it easy to reconstruct. Table 4.2 summarizes the values for each CNN layer used in this study.

Table 4.2. Summary of the values for each CNN layer used in this study.

Layer	Size- in	Size- out	Kernel	Param	FLPS
Conv1	220 X 220 X 3	110 X 110 X 64	7 X 7 X 3,2	9K	115M
Pool1	110 X 110 X 64	55 X 55 X 64	3 X 3 X 64,2	0	
Rnorm1	55 X 55 X 64	55 X 55 X 64		0	
Conv2a	55 X 55 X 64	55 X 55 X 64	1 X1 X64,1	4K	13M
Conv2	55 X 55 X 64	55 X 55 X 192	3 X 3 X64,1	111K	335M
Rnorm2	55 X 55 X 192	55 X 55 X 192		0	
Pool2	55 X 55 X 192	28 X 28 X 192	3 X 3 X 192,2	0	
Conv3a	28 X 28 X 192	28 X 28 X 192	1 X1 X192,1	37K	29M
Conv3	28 X 28 X 192	28 X 28 X 384	3 X 3 X 192,1	664K	521M
Pool3	28 X 28 X 384	14 X 14 X 384	3 X 3 X 192,2	0	
Conv4a	14 X 14 X 384	14 X 14 X 384	1 X1 X384,1	148K	29M
Conv4	14 X 14 X 384	14 X 14 X 256	3 X 3 X 384,1	885K	173M
Conv5a	14 X 14 X 256	14 X 14 X 256	1 X1 X 256,1	66K	13M
Conv5	14 X 14 X 256	14 X 14 X 256	3 X 3 X 256,1	590K	116M
Conv6a	14 X 14 X 256	14 X 14 X 256	1 X1 X 256,1	66K	13M
Conv6	14 X 14 X 256	14 X 14 X 256	3 X 3 X 256,2	590K	116M
Pool4	14 X 14 X 256	7 X 7 X 256		0	
contact	7 X 7 X 256	7 X 7 X 256		0	
Fc1	7 X 7 X 256	1 X 32 X 128	Maxout p=2	103M	103M
Fc2	1 X 32 X 128	1 X 32 X 128	Maxout p=2	34M	34M
Fc7128	1 X 32 X 128	1 X 1 X 128		524K	0.5M
L2	1 X 1 X 128	1 X 1 X 128		0	
Total				140M	1.6B

We investigated architectural changes and preprocessing techniques that could improve the overall performance/efficiency of the models after one or more models have been selected. Using Deepfake models, a Python library designed specifically for this purpose, we were able to experiment with triplet loss trained networks and gain access to pre-trained face reconstruction models via CNN.

Figure 4.12 shows the basic diagram for several DL designs. The initial step involved collecting the data and extracting the relevant features. Eight neural network topologies were employed, and each was evaluated based on five separate evaluation metrics: precision, precision, F1-score, recall, and Receiver Operating Characteristic Curve (ROC curve). This stage extracts several features from the feedback images. Using a filter of a certain size (P*P), the input image is convolved with the filter. Further stages can then use this data to learn more about what they see in the corners and edges of the image. After that, it undergoes a stage of pooling. This stage achieves its primary goal of reducing the size of a convoluted feature map by eliminating the connections between levels and executing autonomously on each element plan. Pooling can be conducted in a variety of ways to achieve distinct

outcomes. The Max Pooling technique is used to perform downsampling (pooling) on a feature map by identifying the maximum value within patches of the feature map. This layer is often used after a convolutional layer. Incorporating a limited degree of translation invariance ensures that most pooled outputs are essentially unaffected by variations in image size [148].

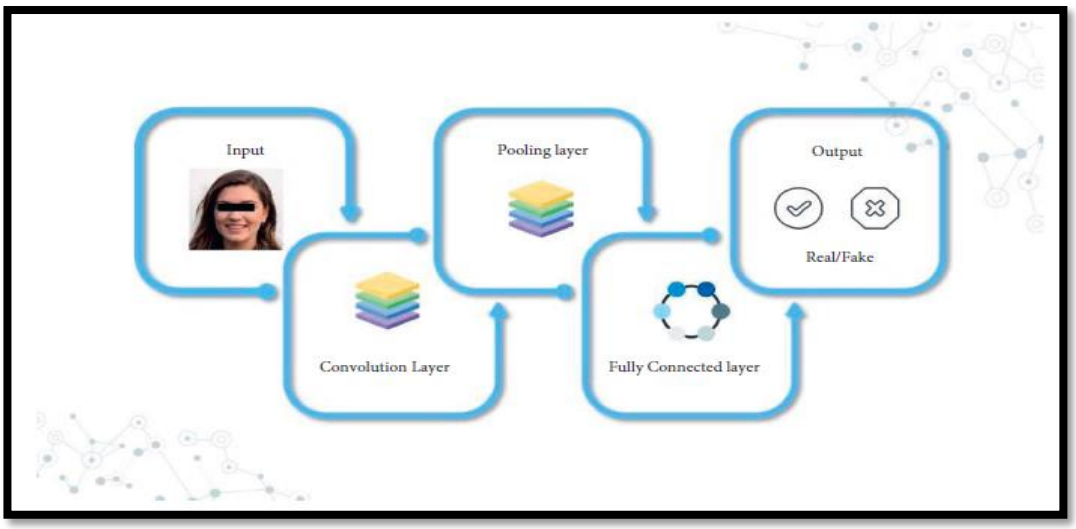

Figure 4.12. The employment method [148].

4.3. PERFORMANCE METRICS

Accuracy: indicates how close the predictive algorithm is to the goal or actual outcome (fake versus real), i.e., the number of times the prototype was able to make accurate predictions out of all the predictions it chose to make. The number of correct predictions and the total number of predictions made represent the model's true prediction and total predictions, respectively see in Eq. (4.1.), [148].

$$Accuracy = \frac{number\ of\ correct\ predictions}{total\ number\ of\ predictions\ made} \quad (4.1)$$

Precision: indicates the stability of the results, regardless of how close they are to the true value while using the target label. Eq. (4.2) defines the ratio that indicates the proportion of correct identifications by model. In Eq. (4.2), TP is the number of true positives, and FP is the number of false positives [148].

$$Precision = \frac{TP}{TP + FP} \tag{4.2}$$

Recall: is the proportion of correct predictions that the model has correctly detected. This ratio is represented by Eq. (4.3), where TP is the number of true positives and FN is the number of false negatives. Recall is the ability of the classifier to identify all positive samples [148].

$$Recall = \frac{TP}{TP + FN} \tag{4.3}$$

F1-score: By balancing precision and recall, the F1 score indicates a model's ability to reliably predict both TP and TN (True Negative) classes. F1-score is the harmonic mean of precision and recall. For Deepfake categorization, the score is a more appropriate metric for evaluating model performance, given both classes are significant and the relative contributions of precision and recall to an F1 score are greater than equal. The equation for the F1-score is given by Eq. (4.4) [148].

$$F1 - score = \frac{2 * (Precision * Recall)}{Precision + Recall} \tag{4.4}$$

PART 5

RESULTS

5.1. DEEPFAKE DETECTION BY USING DFDC DATASET

This thesis presents a research methodology aimed at improving the effectiveness of deepfake models by using CNN with improved logic within MATLAB 2019b. The approach involves incorporating depth information into the CNN model, using either a stereo camera or a depth sensor, to address instances of fake face images, such as those displayed on external screens or printed on paper. In essence, the detection-oriented CNN model will only initiate face reconstruction when it identifies an RGB image with depth information, as opposed to the conventional flat RGB images.

Figures 5.1- 5.9 present experimental findings illustrating how the proposed method use augmented CNN to generate successful triplets by producing faces comparable to those in the DFDC dataset but with different feature vectors and labels. Large datasets are in short supply for firms like Google and Apple to employ in training their state-of-the-art DL models, making this method helpful for researchers. In addition, this approach aims to improve the accuracy of Deepfake identification by developing a system that can effectively categorize Deepfake images while minimizing data loss. Furthermore, it aims to comprehensively explore the potential consequences and future aspects associated with the application of DL methods in the field of deepfake detection.

Figure 5.1. Sample images-1 for evaluation.

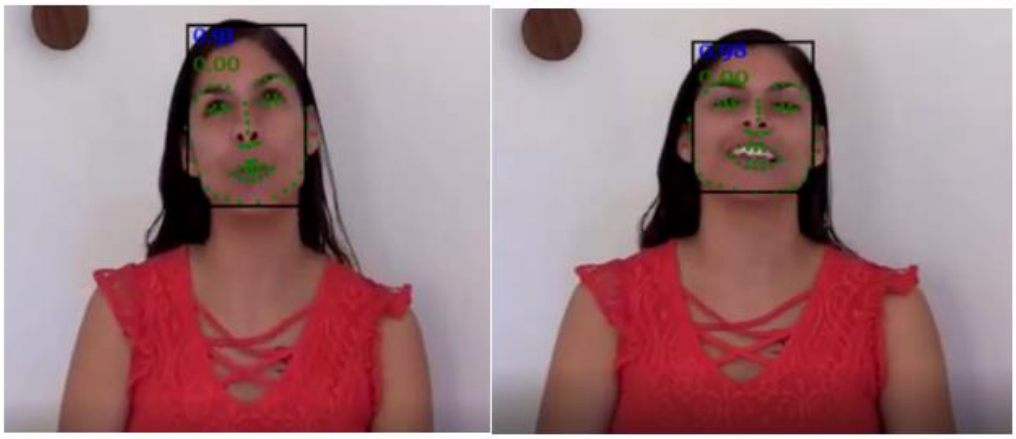

Figure 5.2. Sample video image-1 during training and validation.

Figure 5.3. Initialization image-1 in applied model.

Figure 5.4. Sample image-2 for evaluation.

Figure 5.5. Image sample-2 during training and validation of the Deepfake network.

Figure 5.6. Initialization image-2 in applied model.

Figure 5.7. Video sample images-3 for evaluation.

Figure 5.8. The video image sample-3 during training and validation.

Figure 5.9. Initialization image-3 in applied model.

Figure 5.10 shows 12700 for TPs, 757 for FPs, 589 for FNs, and 12868 for TNs in the confusion matrix for the DFDC dataset, and Table 5.1 compares the results across various models using different evaluation metrics for the DFDC dataset. VGG-Face

achieved a remarkable 99.1% accuracy rate on our training set, outperforming all other pre-trained neural networks. However, the least efficient design, VGG16, achieved an accuracy rate of 93%. DenseNet121 and ResNet50 showed comparable performance, both achieving an accuracy rate of 98%, resulting in a tie for second place. Both the DenseNet201 and DenseNet169 models achieved high accuracy rates of 97% and 96%, respectively. The highest achievable accuracy (99%) was achieved by combining four models. The models used were ResNet50, VGGFace, DenseNet169 and DenseNet121. VGGFace achieved a recall rate of 98%. The DenseNet201 and VGG19 models performed similarly to the VGGFace model, with all three achieving 97% recall. The VGGFace architecture demonstrated the highest F1 score, reaching an impressive 99. The DenseNet121 model was observed to have the lowest F1 score at 85%. ResNet50 achieved an F1 score of 98%, positioning it as the second highest performing model. The DenseNet121 design had the lowest AUC value, while the VGGFace architecture had the highest AUC value, with achieving a score of 99.9%. The RACNN model proposed demonstrated a high level of accuracy, achieving a rate of 95.00% on the DFDC dataset. The custom architecture demonstrated a precision of 94.37% and a recall of 95.57%. Despite achieving a recall rate of 94.37%, the F1 score dropped to 94.97%. In addition, a remarkable AUC of 94% was achieved.

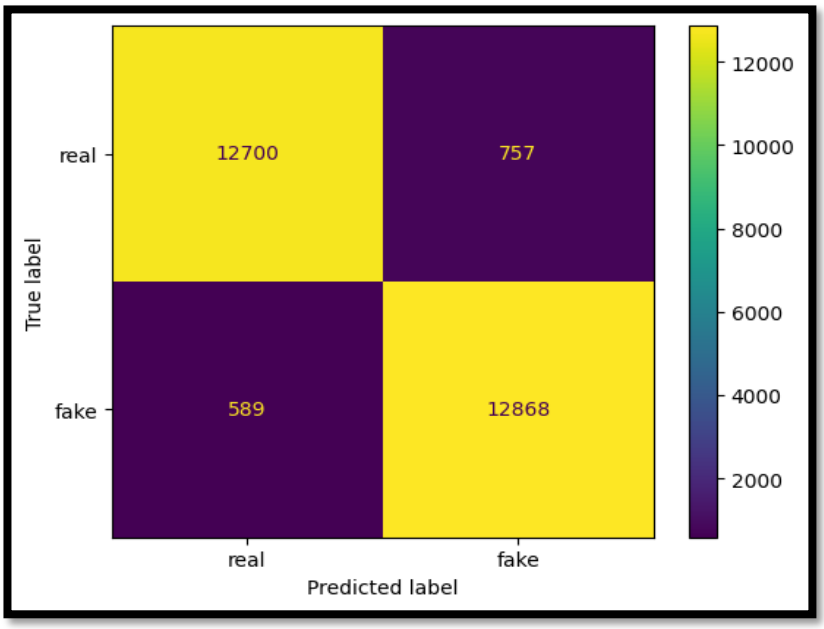

Figure 5.10. Confusion matrix for DFDC dataset.

Table 5.1. Comparison of the many models on DFDC dataset.

Model	Accuracy	Precision	Recall	F1 score	AUC
VGG19	0.940	0.910	0.970	0.940	0.9870
VGG16	0.930	0.930	0.920	0.920	0.9770
VGG-Face	0.991	0.990	0.980	0.990	0.9990
DenseNet169	0.960	0.990	0.920	0.950	0.9960
DenseNet201	0.970	0.960	0.970	0.960	0.9940
DenseNet121	0.980	0.990	0.750	0.850	0.9710
ResNet50	0.980	0.990	0.950	0.980	0.9970
Our Model	0.9500	0.9437	0.9557	0.9497	**0.940**

To demonstrate the effectiveness of our CNN model, we first used several features of different dimensions within a sliding window that spans the entire Deepfake network video. During training, these features act as "weak classifiers" that distinguish whether they contribute to facial attributes. The classifier identifies features that resemble those of human faces, effectively ignoring image regions with minimal or no features that don't contribute to subsequent computations. Using a multi-CNN architecture, CNNs are used to construct the face bounding box. We show how multiple iterations of network propagation can refine many imprecise bounding boxes into a manageable, high-quality selection. Figure 5.11 shows the training and validation accuracies evaluated after 50 training epochs, where an accuracy of 95.00% is achieved.

Figure 5.11. Graph of training and validation accuracy.

In addition, it is important to consider the dataset to be used. The reasoning augmented CNN (labeled faces in the wild) public dataset seems excellent for this purpose, assuming the images are rotated and cropped appropriately, based on the publications cited. Finally, to improve the accuracy and minimize the loss of the model, we undertook the task of adjusting the hyperparameters and using regularization or dropout approaches. This strategy is predicated on the assumption that we lack the capacity to train the model internally. In this scenario, an examination was conducted to explore potential methodologies that could be utilized to enhance the efficiency of the facial reconstruction procedure by leveraging state-of-the-art ML models. Figure 5.12 shows the training and validation loss over 50 training epochs, reaching a minimal value of 0.675.

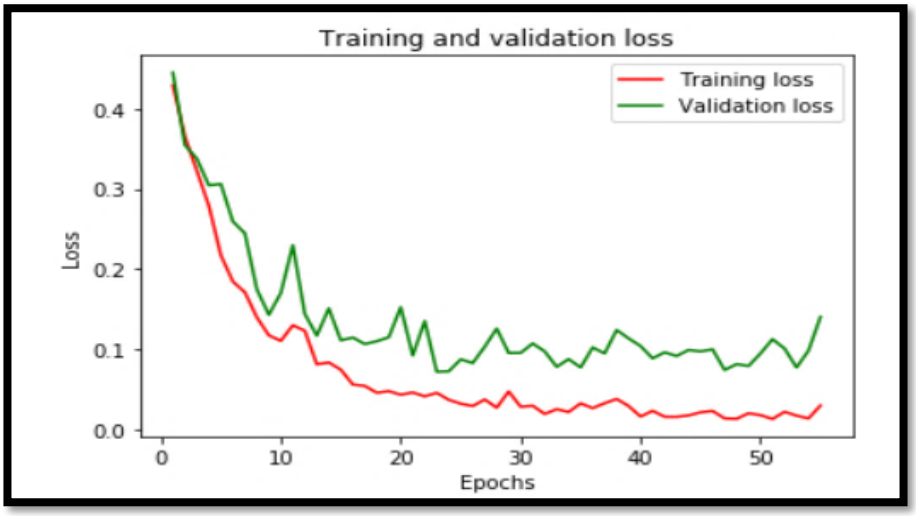

Figure 5.12. Graph of training and validation loss.

Table 5.2 provides a comparative evaluation of our research results in relation to several previous studies. Several studies [131] in this area used a RNN model and achieved a comparable level of accuracy of approximately 54.08%. In another study [132], the use of the Vision Transformer with Xception Network (ViXNet) model resulted in an accuracy rate of 83.18%. Similarly, researchers in another study [133] reported an accuracy rate of 94.43% when using the LSTM model in their research. Meanwhile, another research [134] used a CNN model and achieved an accuracy rate of 84.4%. Finally, the study conducted by a separate group of authors [135], which incorporated the Xception model, demonstrated an accuracy of 81.6%.

Table 5.2. Performance of our model in comparison to existing approaches in terms of different evaluation metrics on DFDC dataset.

Reference	Method / Model	Accuracy	Precision	Recall	AUC
[153]	RNN	54.08%	24.96%	35.08%	51%
[154]	ViXNet	83.18%	-	-	90.26%
[155]	LSTM	94.43%	-	94.30%	-
[156]	CNN	84.40%	-	-	84%
[157]	Xception	81.60%	-	-	66.16%
Our Model	**RACNN + Donald Trump**	**95.00%**	**94.37%**	**95.56%**	**94.00%**

5.2. DEEPFAKE DETECTION BY USING FACEFORENSICS++ DATASET

The methodology includes face alignment, Gaussian Newton optimization, and image blending. Figure 5.13 shows the creation of a manipulated video using a DL technique called CNN Merger with Rationale-Augmented. It transforms an image of Nicolas Cage into a deepfake resembling Donald Trump, highlighting the use of DL algorithms and CNNs in conjunction with the Rationale-Augmented approach to create highly realistic, deceptive video content.

Figure 5.13. The detection of motion in the input images across multiple evaluations of the face.

The experimental results shown in Figures 5.14 and 5.14 demonstrate the feasibility of using an improved CNN model to accurately construct correct triplets using the FaceForencies++ dataset. This dataset generates faces with similar features but different feature vectors and labels. Figure 5.14 illustrates the application of DL models to identify the authenticity of images of Donald Trump and Nicolas Cage. In addition, the proposed technique involves analyzing video frames to identify

anomalies in motion patterns that could potentially serve as indicators of deepfake manipulation. Figure 5.15 shows the motion detection process applied to various instances of character transformation, with a specific focus on Donald Trump.

Figure 5.14. Facial landmarks are detected using a state-of-the-art face alignment network that accurately captures 2D and 3D coordinates.

Figure 5.15. The detection of motion in the input images across multiple evaluations of the face.

Figure 5.16 shows 22156 for true positives, 1514 for false positives, 1327 for false negatives, and 22343 for true negatives in the confusion matrix for the FF++ dataset. The RACNN model was designed considering several factors such as the quality and quantity of the training data and the inclusion of regularization methods. A comparison was made with the CNN model where the collected data was trained and validated for 50 epochs. The RACNN model proposed achieved an accuracy of 94.00% on the FF++ dataset. The custom architecture demonstrated a precision of

93.60% and a recall of 94.35%. Despite achieving a recall rate of 94.35%, the F1 score dropped to 93.97%. In addition, a remarkable AUC of 94% was achieved. The model loss was measured during the training and validation phases for both datasets.

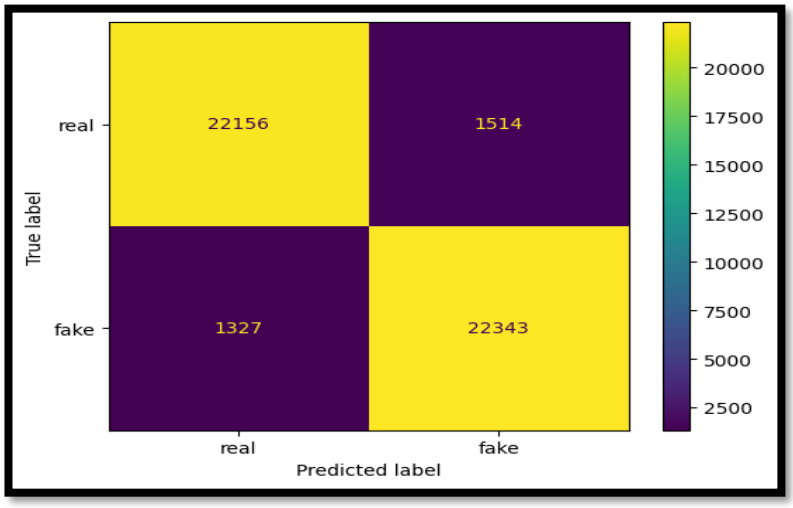

Figure 5.16. Confusion matrix for FF++ dataset.

Table 5.3 provides a detailed evaluation of our model's performance in terms of accuracy relative to existing methods, using the FF++ dataset as the basis for comparison. Several research efforts in Deepfake detection have used different neural network architectures and strategies to achieve accurate results. For example, in [136], the VGG16 and ResNet50 models emerged as robust contenders, with accuracies approaching 86.61% and 75.46%, respectively. A different approach in a separate study [137] focused on the Resnet18 model, which showed a remarkable accuracy rate of 92.23%. In a parallel study [138], researchers harnessed the potential of the AMTENnet model and achieved an admirable accuracy rate of 90.11%. Notably, another scientific investigation [139] turned to the VGG16 model, achieving a commendable accuracy rate of 91.21%, while another investigation [140] ventured into the realm of the EfficientNet model, yielding an accuracy rate of 85.84%. Finally, Efficient Capsule Network [141] achieved a remarkable accuracy rate of 94.51%. These comparative results provide valuable insights into the effectiveness of different deepfake detection methods and contribute significantly to the ongoing discourse and progress in this critical research area.

Table 5.3. Performance of our model in comparison to existing approaches in terms of accuracy on FF++ dataset.

Reference	Model	Accuracy
[158]	VGG16	81.61%
[158]	ResNet50	75.46%
[159]	Resnet18	92.23%
[160]	AMTENnet	90.11%
[161]	VGG16	91.21%
[162]	EfficientNet	85.84%
[163]	Efficient-Capsule Network	94.51%
Our model	RACNN + Donald Trump	94.00%

These results underscore the effectiveness of the RACNN in accurately distinguishing between real and fake images, demonstrating its potential utility in real-world scenarios. The remarkable accuracy of the model is due to its unique design, which incorporates advanced architectural techniques such as sophisticated CNN layers, attention mechanisms, and the integration of temporal information. This holistic approach enables the RACNN to capture intricate features and patterns associated with real and fake images, setting it apart from other models.

PART 6

CONCLUSION

A Deepfake network architecture has significant potential in the field of global security research, providing solutions to address misbehavior and security challenges through facial reconstruction. Our Deepfake architecture might be applied to various applications, ranging from security to unfamiliar observational contexts such as face detection and reconstruction. In this thesis, we evaluated two datasets, the first dataset DFDC dataset using a MATLAB R2019a-based, Rationale-Augmented Convolutional Neural Network (RACNN), and the second FaceForensics++ dataset. The CNN strategy is practically constant, and there is little variation in computational cost between the two methods. When applying the Donald Trump filter to the Deepfake video, we found that a low computational cost was necessary to establish a faster link based on the association between the faces. This large dataset has been produced many times, making it perfect for precise grouping and splitting. Moreover, the simple implementation of the CNN model facilitated its integration with a partitioning technique, yielding a remarkable accuracy of 94.9989% for the DFDC dataset and 93.9987% for the FaceForensics++ dataset.

Future research may explore unsupervised assembly techniques such as autoencoders to evaluate their effectiveness in Deepfake classification within CNN algorithms. Recommendations for future studies are further outlined below:

- Classification methods could be developed to analyze and then flag users of social networking sites who upload pictures or videos before being published online. This would help stop the spread of false information and prevent its further dissemination.
- We intend to enhance further the performance of DL algorithms and investigate the implementation of video steganography, steganalysis, and

cryptanalysis in identifying and classifying honest and persuasive facial images. This would allow us to classify and identify real and fake face images.

- Generate training data that can be used to improve model effectiveness by collecting and testing a variety of hidden classifiers.
- Implement the patch-dependent fuzzy rough set feature selection technique to identify deepfakes through anomaly detection in originally used patches.
- Combining local image techniques with inter-plus ensemble modeling approaches, such as holistic, content, noise level, and steganographic, to achieve improved performance by considering different aspects of image characteristics.
- Evaluating the performance of EffectiveNet on the Deepfake image datasets used in this thesis to identify potentially suitable modeling frameworks for ensemble-based approaches.

REFERENCES

1. A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, “FaceForensics++: Learning to Detect Manipulated Facial Images,” ***2019 IEEE/CVF International Conference on Computer Vision (ICCV),*** Oct. 2019.

2. E. M. Torralba, “Fibonacci Numbers as Hyperparameters for Image Dimension of a Convolu-tional Neural Network Image Prognosis Classification Model of COVID X-ray Images,” ***International Journal of Multidisciplinary: Applied Business and Education Research***, vol. 3, no. 9, pp. 1703–1716, Sep. 2022.

3. Y. Li, M.-C. Chang, and S. Lyu, “In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking,” ***2018 IEEE International Workshop on Information Forensics and Security (WIFS)***, Dec. 2018.

4. P. Korshunov and S. Marcel, “Vulnerability assessment and detection of Deepfake videos,” ***2019 International Conference on Biometrics (ICB)***, Jun. 2019.

5. J. Fridrich and J. Kodovsky, “Rich Models for Steganalysis of Digital Images,” ***IEEE Transactions on Information Forensics and Security***, vol. 7, no. 3, pp. 868–882, Jun. 2012.

6. R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, “Deepfakes and beyond: A Survey of face manipulation and fake detection,” ***Information Fusion***, vol. 64, pp. 131–148, Dec. 2020.

7. P. Yu, Z. Xia, J. Fei, and Y. Lu, “A Survey on Deepfake Video Detection,” ***IET Biometrics***, vol. 10, no. 6, pp. 607–624, Apr. 2021.

8. W. M. Wubet*, “The Deepfake Challenges and Deepfake Video Detection,” ***International Journal of Innovative Technology and Exploring Engineering***, vol. 9, no. 6, pp. 789–796, Apr. 2020.

9. D. Guera and E. J. Delp, “Deepfake Video Detection Using Recurrent Neural Networks,” ***2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)***, Nov. 2018.

10. E. Sabir, et al. Recurrent convolutional strategies for face manipulation detection in videos. ***Interfaces (GUI)***, 3.1: 80-87, 2019.

11. O. de Lima, S. Franklin, S. Basu, B. Karwoski and A. George, "Deepfake detection using spatiotemporal convolutional networks", ***IEEE Access***, 2020.

12. Z. Tianyu, M. Zhenjiang, and Z. Jianhu, “Combining CNN with Hand-Crafted Features for Image Classification,” ***2018 14th IEEE International Conference on Signal Processing (ICSP)***, Aug. 2018.

13. S. Das, S. Seferbekov, A. Datta, M. S. Islam and M. R. Amin, "Towards Solving the DeepFake Problem: An Analysis on Improving DeepFake Detection using Dynamic Face Augmentation," ***2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)***, Montreal, BC, Canada, 2021.

14. DeepFaceLab. https://github.com/iperov/DeepFaceLab, accessed: 2020-03-30.

15. faceswap. https://github.com/deepfakes/faceswap, accessed: 2020-03-30.

16. MarekKowalski/FaceSwap. https://github.com/MarekKowalski/ FaceSwap, accessed: 2020-04-06.

17. C.-K. Lee, Y.-J. Cheon, and W.-Y. Hwang, “Least Squares Generative Adversarial Networks-Based Anomaly Detection,” ***IEEE Access***, vol. 10, pp. 26920–26930, 2022.

18. Y. Que and H. J. Lee, “Densely Connected Convolutional Networks for Multi-Exposure Fusion,” ***2018 International Conference on Computational Science and Computational Intelligence (CSCI)***, Dec. 2018.

19. Y. Sun, X. Wang, and X. Tang, “Deep Learning Face Representation from Predicting 10,000 Classes,” ***2014 IEEE Conference on Computer Vision and Pattern Recognition***, Jun. 2014.

20. A. Kolekar and V. Dalal, “Barcode Detection and Classification using SSD (Single Shot Multibox Detector) Deep Learning Algorithm,” ***SSRN Electronic Journal***, 2020.

21. C. Xiaopeng, C. Jiangzhong, L. Yuqin, and D. Qingyun, “Improved Training of Spectral Normalization Generative Adversarial Networks,” 2020 2nd World Symposium on Artificial Intelligence (WSAI), Jun. 2020.

22. O. Russakovsky et al., “ImageNet Large Scale Visual Recognition Challenge,” ***International Journal of Computer Vision***, vol. 115, no. 3, pp. 211–252, Apr. 2015.

23. H. Mo, B. Chen, and W. Luo, “Fake Faces Identification via Convolutional Neural Network,” ***Proceedings of the 6th ACM Workshop on Information Hiding and Multimedia Security***, Jun. 2018.

24. L. Dang, S. Hassan, S. Im, J. Lee, S. Lee, and H. Moon, “Deep Learning Based Computer-Generated Face Identification Using Convolutional Neural Network,” ***Applied Sciences***, vol. 8, no. 12, p. 2610, Dec. 2018.

25. Z. Liu, Y. Niu, and Q. Qu, “Fingerprint Identification using Ridge Lines,” 2022 ***3rd International Conference on Computer Vision, Image and Deep Learning & International Conference on Computer Engineering and Applications (CVIDL & ICCEA)***, May 2022

26. C.-C. Hsu, Y.-X. Zhuang, and C.-Y. Lee, “Deep Fake Image Detection Based on Pairwise Learning,” ***Applied Sciences***, vol. 10, no. 1, p. 370, Jan. 2020.

27. S. Unnikrishnan and A. Eshack, “Face spoof detection using image distortion analysis and image quality assessment,” ***2016 International Conference on Emerging Technological Trends (ICETT)***, Oct. 2016.

28. I. Korshunova, W. Shi, J. Dambre, and L. Theis, “Fast Face-Swap Using Convolutional Neural Networks,” ***2017 IEEE International Conference on Computer Vision (ICCV),*** Oct. 2017.

29. Y. Zhang, L. Zheng, and V. L. L. Thing, “Automated face swapping and its detection,” ***2017 IEEE 2nd International Conference on Signal and Image Processing (ICSIP),*** Aug. 2017.

30. X. Wang, N. Thome, and M. Cord, “Gaze latent support vector machine for image classification improved by weakly supervised region selection,” ***Pattern Recognition***, vol. 72, pp. 59–71, Dec. 2017.

31. S. Bai, “Growing random forest on deep convolutional neural networks for scene categorization,” ***Expert Systems with Applications***, vol. 71, pp. 279–287, Apr. 2017.

32. L. Zheng, S. Duffner, K. Idrissi, C. Garcia, and A. Baskurt, “Siamese multi-layer perceptrons for dimensionality reduction and face identification,” ***Multimedia Tools and Applications***, vol. 75, no. 9, pp. 5055–5073, Aug. 2015.

33. X. Xuan, B. Peng, W. Wang, and J. Dong, “On the Generalization of GAN Image Forensics,” ***Lecture Notes in Computer Science***, pp. 134–141, 2019.

34. P. Yang, R. Ni, and Y. Zhao, “Recapture Image Forensics Based on Laplacian Convolutional Neural Networks,” ***Lecture Notes in Computer Science***, pp. 119–128, 2017.

35. B. Bayar and M. C. Stamm, “A Deep Learning Approach to Universal Image Manipulation Detection Using a New Convolutional Layer,” ***Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security***, Jun. 2016.

36. X. Liu, X. Wang, and S. Matwin, “Interpretable Deep Convolutional Neural Networks via Meta-learning,” ***2018 International Joint Conference on Neural Networks (IJCNN)***, Jul. 2018.

37. Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, Matthias Nießner. "FaceForensics++: Learning to Detect Manipulated Facial Images." ***In IEEE International Conference on Computer Vision Workshops (ICCVW)***, 2019. DOI: 10.1109/ICCVW.2019.00173

38. Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Nießner, Patrick Pérez, Christian Richardt, Michael Zollhöfer. "Deep Video Portraits." ***In ACM Transactions on Graphics (TOG)***, 2018. DOI: 10.1145/3197517.3201283

39. Zitong Yu, Chenxu Zhao, Zhengming Ding, Mingming Jiang, Jinqiao Wang, Xiaopeng Hong. "Learning Deep Models for Face Anti-Spoofing: Binary or Auxiliary Supervision." ***In IEEE Transactions on Information Forensics and Security***, 2019. DOI: 10.1109/TIFS.2018.2889071

40. Xin Yang, Yuezun Li, Zheng Wang, Yuan Yuan, Chenggang Yan. "Face X-ray for More General Face Forgery Detection." ***In IEEE Transactions on Information Forensics and Security***, 2020. DOI: 10.1109/TIFS.2019.2927976

41. Muhammad Usama, Mian Ahsan Iftikhar, Muhammad Arsalan. "Detecting Deepfake Videos from the Lip Movements Using CNN." ***In IEEE Access***, 2020. DOI: 10.1109/ACCESS.2020.3011343

42. C. Zhang, Y. Feng, B. Qiang, and J. Shang, “Wasserstein Generative Recurrent Adversarial Networks for Image Generating,” ***2018 24th International Conference on Pattern Recognition (ICPR)***, Aug. 2018.

43. K. Lei, M. Mardani, J. M. Pauly, and S. S. Vasanawala, “Wasserstein GANs for MR Imaging: From Paired to Unpaired Training,” ***IEEE Transactions on Medical Imaging***, vol. 40, no. 1, pp. 105–115, Jan. 2021.

44. C.-K. Lee, Y.-J. Cheon, and W.-Y. Hwang, “Least Squares Generative Adversarial Networks-Based Anomaly Detection,” ***IEEE Access***, vol. 10, pp. 26920–26930, 2022.

45. G. Keren, J. Deng, J. Pohjalainen, and B. Schuller, “Convolutional Neural Networks with Data Augmentation for Classifying Speakers’ Native Language,” ***Interspeech 2016***, Sep. 2016.

46. I. J. Goodfellow, J. P. Abadie, M. Mirza et al., "Generative adversarial nets, "NIPS" 14," ***Proceedings of the 27th Inter- national Conference on Neural Information Processing Sys- tems***, vol. 2, pp. 2672–2680, 2014.

47. K. Kulkarni, “DeepFake Detection: A survey of countering malicious DeepFakes,” ***International Journal for Research in Applied Science and Engineering Technology***, vol. 10, no. 6, pp. 4492–4495, Jun. 2022.

48. T. Jung, S. Kim, as well as K. Kim, "DeepVision: Deep fakes detection utilization human eye blinking pattern," ***IEEE Access***, vol. 8, pp. 83144–83154, 2020.

49. M. Westerlund, "The emergence of Deepfake technology: a review," ***Technology Innovation Management Review***, vol. 9, no. 11, pp. 39–52, 2019.

50. M.-H. Maras as well as A. Alexandrou, "Determining authenticity of video evidence in the age of artificial intelligence as well as in the wake of Deepfake videos," ***International Journal of Evidence as well as Proof***, vol. 23, no. 3, pp. 255–262, 2019.

51. A. M. Almars, "Deep fakes detection techniques utilization deep learning: a survey," ***Journal of Computer as well as Communications***, vol. 9, no. 5, pp. 20–35, 2021.

52. L. Guarnera, O. Giudice, as well as S. Battiato, "DeepFake detection by analyzing convolutional traces," ***in Proceedings of the 2020 IEEE/CVF Conference on Computer Vision as well as Pattern Rec- ognition Workshops (CVPRW)***, pp. 2841–2850, Seattle, WA, U.S.A., 2020.

53. A. Punnappurath as well as M. S. Brown, "Learning raw image reconstruction-aware deep image compressors," ***IEEE Transactions on Pattern Analysis as well as Machine Intelligence***, vol. 42, no. 4, pp. 1013–1019, 2020.

54. Z. Cheng, H. Sun, M. Takeuchi, as well as J. Katto, "Energy compaction-based image compression utilization convolutional AutoEncoder," ***IEEE Transactions on Multimedia***, vol. 22, no. 4, pp. 860–873, 2020.

55. J. Chorowski, R. J. Weiss, S. Bengio, as well as A. van den Oord, "Unsupervised speech representation learning utilization WaveNet autoencoders," ***IEEE/ACM Transactions on Audio, Speech, as well as Language Processing***, vol. 27, no. 12, pp. 2041–2053, 2019.

56. Faceswap, "Deep fakes software for all," ***https://github.com/ Deep-fakes/faceswap***.

57. FakeApp 2.2.0, https://www.malavida.com/en/soft/fakeapp/.

58. DeepFaketf, "Deepfake based on tensorflow," https://github.com/StromWine/DeepFake%20tf.

59. DFaker, https://github.com/dfaker/df.

60. DeepFaceLab, https://github.com/iperov/DeepFaceLab.
61. Faceswap-GAN, https://github.com/shaoanlu/faceswap-GAN.

62. Keras-VGGFace, "VGGFace implementation with Keras frame-work," https://github.com/rcmalli/keras-vggface.

63. FaceNet, https://github.com/davidsandberg/facenet.

64. CycleGAN, https://github.com/junyanz/pytorch-CycleGAN- and-pix2pix.

65. R. Chesney and D. K. Citron, “Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security,” ***SSRN Electronic Journal***, 2018.

66. T. T. Nguyen et al., “Deep Learning for Deepfakes Creation and Detection: A Survey,” ***SSRN Electronic Journal***, 2022.

67. R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, “Deepfakes and beyond: A Survey of face manipulation and fake detection,” ***Information Fusion***, vol. 64, pp. 131–148, Dec. 2020.

68. A. Kumar, A. Bhavsar, and R. Verma, “Detecting Deepfakes with Metric Learning,” ***2020 8th International Workshop on Biometrics and Forensics (IWBF)***, Apr. 2020.

69. Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, “Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics,” ***2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)***, Jun. 2020.

70. S. Lyu, “Deepfake Detection: Current Challenges and Next Steps,” ***2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)***, Jul. 2020

71. M. Boháček and H. Farid, “Protecting world leaders against deep fakes using facial, gestural, and vocal mannerisms,” ***Proceedings of the National Academy of Sciences***, vol. 119, no. 48, Nov. 2022.

72. F. Matern, C. Riess, and M. Stamminger, “Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations,” ***2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)***, Jan. 2019.

73. F. Matern, C. Riess, and M. Stamminger, “Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations,” ***2019 IEEE Winter Applications of Computer Vision Workshops (WACVW),*** Jan. 2019.

74. X. Yang, Y. Li, and S. Lyu, “Exposing Deep Fakes Using Inconsistent Head Poses,” ***ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)***, May 2019

75. LI, Yuezun; LYU, Siwei. Exposing deepfake videos by detecting face warping artifacts. ***arXiv preprint arXiv:1811.00656***, 2018.

76. Y. Li, M.-C. Chang, and S. Lyu, "In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking," ***2018 IEEE International Workshop on Information Forensics and Security (WIFS)***, Dec. 2018.

77. T. Jung, S. Kim, and K. Kim, "DeepVision: Deepfakes Detection Using Human Eye Blinking Pattern," ***IEEE Access***, vol. 8, pp. 83144–83154, 2020.

78. D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: a Compact Facial Video Forgery Detection Network," ***2018 IEEE International Workshop on Information Forensics and Security (WIFS)***, Dec. 2018.

79. P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, "Two-Stream Neural Networks for Tampered Face Detection," ***2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW),*** Jul. 2017.

80. M. Goljan and J. Fridrich, "CFA-aware features for steganalysis of color images," ***Media Watermarking, Security, and Forensics 2015***, Mar. 2015.

81. A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "FaceForensics++: Learning to Detect Manipulated Facial Images," ***2019 IEEE/CVF International Conference on Computer Vision (ICCV)***, Oct. 2019.

82. H. H. Nguyen, J. Yamagishi, and I. Echizen, "Capsule-forensics: Using Capsule Networks to Detect Forged Images and Videos," ***ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)***, May 2019.

83. SIMONYAN, Karen; ZISSERMAN, Andrew. Very deep convolutional networks for large-scale image recognition. ***arXiv preprint arXiv***:1409.1556, 2014.

84. D. Guera and E. J. Delp, "Deepfake Video Detection Using Recurrent Neural Networks," ***2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)***, Nov. 2018.

85. DE LIMA, Oscar, et al. Deepfake detection using spatiotemporal convolutional networks. ***arXiv preprint arXiv:2006.14749***, 2020.

86. Y. Wang and A. Dantcheva, "A video is worth more than 1000 lies. Comparing 3DCNN approaches for detecting deepfakes," ***2020 15th IEEE International***

Conference on Automatic Face and Gesture Recognition (FG 2020), Nov. 2020.

87. K. Hara, H. Kataoka, and Y. Satoh, "Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet?," ***2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Jun. 2018.***

88. A. Traore and M. A. Akhloufi, "Violence Detection in Videos using Deep Recurrent and Convolutional Neural Networks," ***2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)***, Oct. 2020.

89. Z. Tianyu, M. Zhenjiang, and Z. Jianhu, "Combining CNN with Hand-Crafted Features for Image Classification," ***2018 14th IEEE International Conference on Signal Processing (ICSP)***, Aug. 2018.

90. A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies and M. Nießner, "Faceforensics: A large-scale video dataset for forgery detection in human faces", ***arXiv:1803.09179***, 2018.

91. D. Guera and E. J. Delp, "Deepfake Video Detection Using Recurrent Neural Networks," ***2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)***, Nov. 2018.

92. S. Agarwal, H. Farid, T. El-Gaaly, and S.-N. Lim, "Detecting Deep-Fake Videos from Appearance and Behavior," ***2020 IEEE International Workshop on Information Forensics and Security (WIFS)***, Dec. 2020.

93. H. H. Nguyen, J. Yamagishi, and I. Echizen, "Capsule-forensics: Using Capsule Networks to Detect Forged Images and Videos," ***ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)***, May 2019.

94. F. Marcon, C. Pasquini, and G. Boato, "Detection of Manipulated Face Videos over Social Networks: A Large-Scale Study," ***Journal of Imaging***, vol. 7, no. 10, p. 193, Sep. 2021.

95. N. Bonettini, E. D. Cannas, S. Mandelli, L. Bondi, P. Bestagini, and S. Tubaro, "Video Face Manipulation Detection Through Ensemble of CNNs," ***2020 25th International Conference on Pattern Recognition (ICPR)***, Jan. 2021.

96. G. Fox, W. Liu, H. Kim, H.-P. Seidel, M. Elgharib, and C. Theobalt, "Videoforensicshq: Detecting High-Quality Manipulated Face Videos," ***2021 IEEE International Conference on Multimedia and Expo (ICME)***, Jul. 2021.

97. Z. Xu et al., “Detecting facial manipulated videos based on set convolutional neural networks,” ***Journal of Visual Communication and Image Representation***, vol. 77, p. 103119, May 2021.

98. E. Sabir, J. Cheng, A. Jaiswal, W. Abd-Almageed, I. Masi and P. Natarajan, "Recurrent convolutional strategies for face manipulation detection in videos", Proc. ***CVPR Workshops***, pp. 80-87, 2019.

99. Y. Wang and A. Dantcheva, “A video is worth more than 1000 lies. Comparing 3DCNN approaches for detecting deepfakes,” ***2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)***, Nov. 2020.

100. H. H. Nguyen, F. Fang, J. Yamagishi, and I. Echizen, “Multi-task Learning for Detecting and Segmenting Manipulated Facial Images and Videos,” ***2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS)***, Sep. 2019.

101. A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, “FaceForensics++: Learning to Detect Manipulated Facial Images,” ***2019 IEEE/CVF International Conference on Computer Vision (ICCV)***, Oct. 2019.

102. D. Xie, P. Chatterjee, Z. Liu, K. Roy, and E. Kossi, “DeepFake Detection on Publicly Available Datasets using Modified AlexNet,” ***2020 IEEE Symposium Series on Computational Intelligence (SSCI)***, Dec. 2020.

103. D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, “MesoNet: a Compact Facial Video Forgery Detection Network,” ***2018 IEEE International Workshop on Information Forensics and Security (WIFS)***, Dec. 2018.

104. Z. Guo, G. Yang, J. Chen, and X. Sun, “Fake face detection via adaptive manipulation traces extraction network,” ***Computer Vision and Image Understanding***, vol. 204, p. 103170, Mar. 2021.

105. I. Amerini, L. Galteri, R. Caldelli, and A. Del Bimbo, “Deepfake Video Detection through Optical Flow Based CNN,” 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), Oct. 2019.

106. M. Bonomi, C. Pasquini, and G. Boato, “Dynamic texture analysis for detecting fake faces in video sequences,” ***Journal of Visual Communication and Image Representation***, vol. 79, p. 103239, Aug. 2021.

107. H. Khalid and S. S. Woo, “OC-FakeDect: Classifying Deepfakes Using One-class Variational Autoencoder,” 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Jun. 2020.

108. M. Kim, S. Tariq, and S. S. Woo, “FReTAL: Generalizing Deepfake Detection using Knowledge Distillation and Representation Learning,” ***2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW),*** Jun. 2021.

109. J. Li, H. Xie, J. Li, Z. Wang, and Y. Zhang, “Frequency-aware Discriminative Feature Learning Supervised by Single-Center Loss for Face Forgery Detection,” ***2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),*** Jun. 2021.

110. H. Liu et al., “Spatial-Phase Shallow Learning: Rethinking Face Forgery Detection in Frequency Domain,” ***2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),*** Jun. 2021.

111. H. H. Nguyen, J. Yamagishi, and I. Echizen, “Capsule-forensics: Using Capsule Networks to Detect Forged Images and Videos,” ***ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)***, May 2019.

112. X. Fan and T. Tjahjadi, “Fusing dynamic deep learned features and handcrafted features for facial expression recognition,” ***Journal of Visual Communication and Image Representation***, vol. 65, p. 102659, Dec. 2019.

113. Isabella di Lenardo, Simone Bianco, Paolo Napoletano, Raimondo Schettini. "ForgeryNet: A Large-Scale Dataset for Forgery Detection in Art." ***In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops***, 2020.

114. J. K. Lewis et al., “Deepfake Video Detection Based on Spatial, Spectral, and Temporal Inconsistencies Using Multimodal Deep Learning,” ***2020 IEEE Applied Imagery Pattern Recognition Workshop (AIPR),*** Oct. 2020

115. Y. Zhang, L. Zheng, and V. L. L. Thing, “Automated face swapping and its detection,” ***2017 IEEE 2nd International Conference on Signal and Image Processing (ICSIP)***, Aug. 2017.

116. Passos LA, Jodas D, da Costa KA, et al (2022) A review of deep learning-based approaches for deepfake content detection. ***arXiv preprint arXiv***:220206095

117. TIWARI, Aniruddha; DAVE, Rushit; VANAMALA, Mounika. Leveraging Deep Learning Approaches for Deepfake Detection: A Review. ***arXiv preprint arXiv:2304.01908***, 2023.

118. T. Karras, T. Aila, S. Laine and J. Lehtinen, "Progressive growing of GANs for improved quality stability and variation", ***Proc. Int. Conf. Learn. Representations***, 2018.

119. Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep Learning Face Attributes in the Wild," ***2015 IEEE International Conference on Computer Vision (ICCV)***, Dec. 2015.

120. T. Karras, S. Laine, and T. Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks," ***2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)***, Jun. 2019.

121. X. Huang and S. Belongie, "Arbitrary Style Transfer in Real-Time with Adaptive Instance Normalization," ***2017 IEEE International Conference on Computer Vision (ICCV)***, Oct. 2017.

122. T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and Improving the Image Quality of StyleGAN," **2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**, Jun. 2020

123. T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen and T. Aila, "Training generative adversarial networks with limited data", ***Proc. Conf. Neural Inf. Process. Syst.***, pp. 12104-12114, 2020

124. S. Liu, P. Cui, W. Zhu, and S. Yang, "Learning Socially Embedded Visual Representation from Scratch," ***Proceedings of the 23rd ACM international conference on Multimedia***, Oct. 2015.

125. Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A Dataset for Recognising Faces across Pose and Age," ***2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)***, May 2018.

126. A. Nech and I. Kemelmacher-Shlizerman, "Level Playing Field for Million Scale Face Recognition," ***2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)***, Jul. 2017

127. J. C. Neves, R. Tolosana, R. Vera-Rodriguez, V. Lopes, H. Proenca, and J. Fierrez, "GANprintR: Improved Fakes and Evaluation of the State of the Art in Face Manipulation Detection," ***IEEE Journal of Selected Topics in Signal Processing***, vol. 14, no. 5, pp. 1038–1048, Aug. 2020

128. Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, “Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics,” ***2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)***, Jun. 2020

129. Y. Mirsky and W. Lee, "The creation and detection of deepfakes: A survey", ***arXiv:2004.11138***, 2020

130. Y. Li, M.-C. Chang, and S. Lyu, “In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking,” 2018 IEEE International Workshop on Information Forensics and Security (WIFS), Dec. 2018

131. B. Dolhansky, R. Howes, B. Pflaum, N. Baram and C. C. Ferrer, "The deepfake detection challenge (DFDC) preview dataset", arXiv:1910.08854v2, 2019

132. Y. Li, M.-C. Chang, and S. Lyu, “In Ictu Oculi: Exposing AI Created Fake Videos by Detecting Eye Blinking,” ***2018 IEEE International Workshop on Information Forensics and Security (WIFS),*** Dec. 2018
133. R. Skibba, “Accuracy Eludes Competitors in Facebook Deepfake Detection Challenge,” ***Engineering***, vol. 6, no. 12, pp. 1339–1340, Dec. 2020

134. P. Korshunov and S. Marcel, "DeepFakes: A new threat to face recognition? Assessment and detection", ***arXiv:1812.08685v1***, 2018

135. C. Sanderson and B. C. Lovell, “Multi-Region Probabilistic Histograms for Robust and Scalable Identity Inference,” ***Lecture Notes in Computer Science***, pp. 199–208, 2009.

136. J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks,” ***2017 IEEE International Conference on Computer Vision (ICCV)***, Oct. 2017

137. F. Schroff, D. Kalenichenko, and J. Philbin, “FaceNet: A unified embedding for face recognition and clustering***,” 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)***, Jun. 2015

138. K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, “Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks,” ***IEEE Signal Processing Letters***, vol. 23, no. 10, pp. 1499–1503, Oct. 2016

139. U. Scherhag, C. Rathgeb, J. Merkle, R. Breithaupt, and C. Busch, “Face Recognition Systems Under Morphing Attacks: A Survey,” ***IEEE Access***, vol. 7, pp. 23012–23026, 2019

140. K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, “Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks,” ***IEEE Signal Processing Letters***, vol. 23, no. 10, pp. 1499–1503, Oct. 2016

141. H. Zhang, S. Venkatesh, R. Ramachandra, K. Raja, N. Damer, and C. Busch, “MIPGAN—Generating Strong and High Quality Morphing Attacks Using Identity Prior Driven GAN,” ***IEEE Transactions on Biometrics, Behavior, and Identity Science***, vol. 3, no. 3, pp. 365–383, Jul. 2021.

142. Y. Zhang et al., “Deepfake Detection System for the ADD Challenge Track 3.2 Based on Score Fusion,” ***Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia***, Oct. 2022

143. H. Sharma and N. Kanwal, “Video interframe forgery detection: Classification, technique & new dataset,” ***Journal of Computer Security***, vol. 29, no. 5, pp. 531–550, Aug. 2021

144. E. Gonzalez-Sosa, J. Fierrez, R. Vera-Rodriguez, and F. Alonso-Fernandez, “Facial Soft Biometrics for Recognition in the Wild: Recent Works, Annotation, and COTS Evaluation,” ***IEEE Transactions on Information Forensics and Security***, vol. 13, no. 8, pp. 2001–2014, Aug. 2018.

145. Available Online: https://www.kaggle.com/c/deepfake-detection-challenge/data.

146. A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J.)ies, and M. Nießner, “Faceforensics++: learning to detect manipulated facial images,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1–11, Seoul, Republic of Korea, 2019.

147. Classification, “ROC curve and AUC,” https://developers. google.com/machine-learning/crash-course/classification/ roc-and-auc.

148. H. S. Shad et al., “Comparative Analysis of Deepfake Image Detection Method Using Convolutional Neural Network,” ***Computational Intelligence and Neuroscience***, vol. 2021, pp. 1–18, Dec. 2021, doi: 10.1155/2021/3111676.

149. G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” ***in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition***, pp. 4700–4708, Honolulu, HI, USA, July 2017

150. https://www.kaggle.com/keras/resnet50.

151. https://www.kaggle.com/shivamb/cnn-architectures-vgg-resnet -inception-tl.

152. https://sefiks.com/2018/08/06/deep-face-recognition-withkeras/.

153. P. Saikia, D. Dholaria, P. Yadav, V. Patel, and M. Roy, “A Hybrid CNN-LSTM model for Video Deepfake Detection by Leveraging Optical Flow Features,” ***2022 International Joint Conference on Neural Networks (IJCNN)***, Jul. 2022.

154. S. Ganguly, A. Ganguly, S. Mohiuddin, S. Malakar, and R. Sarkar, “ViXNet: Vision Transformer with Xception Network for deepfakes based video and image forgery detection,” Expert Systems with Applications, vol. 210, p. 118423, Dec. 2022.

155. I. Masi, A. Killekar, R. M. Mascarenhas, S. P. Gurudatt, and W. AbdAlmageed, “Two-Branch Recurrent Network for Isolating Deepfakes in Videos,” Lecture Notes in Computer Science, pp. 667–684, 2020.

156. Mittal, T., Bhattacharya, U., Chandra, R., Bera, A., & Manocha, D. (2020, October). Emotions don't lie: An audio-visual deepfake detection method using affective cues. ***In Proceedings of the 28th ACM international conference on multimedia*** (pp. 2823-2832).

157. H. Liu et al., “Spatial-Phase Shallow Learning: Rethinking Face Forgery Detection in Frequency Domain,” ***2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)***, Jun. 2021.

158. I. Amerini, L. Galteri, R. Caldelli and A. Del Bimbo, "Deepfake Video Detection through Optical Flow Based CNN," ***2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), Seoul, Korea (South),*** 2019, pp. 1205-1207, doi: 10.1109/ICCVW.2019.00152.

159. S. Aneja and M. Nießner, "Generalized zero and few-shot transfer for facial forgery detection", ***arXiv preprint***, 2020.

160. Z. Guo, G. Yang, J. Chen and X. Sun, "Fake face detection via adaptive manipulation traces extraction network", ***arXiv:2005.04945***, 2020.

161. P. Saikia, D. Dholaria, P. Yadav, V. Patel and M. Roy, "A Hybrid CNN-LSTM model for Video Deepfake Detection by Leveraging Optical Flow Features," ***2022 International Joint Conference on Neural Networks (IJCNN), Padua, Italy***, 2022, pp. 1-7, doi: 10.1109/IJCNN55064.2022.9892905.

162. S. Suratkar, S. Bhiungade, J. Pitale, et al. “Deep-fake video detection approaches using convolutional–recurrent neural networks”. ***Journal of Control and Decision***, pp1–17, 2022.

163. H. Ilyas, A. Javed, K. Malik, et al. “E-Cap Net: an efficient-capsule network for shallow and deepfakes forgery detection”. ***Multimedia Systems*** 29, 2165–2180 (2023).

RESUME

Saadaldeen Rashid AHMED graduated first and elementary education in salah-aldeen city. He completed high school education at (Saad abn abe wakas) high school in Salahaldin. He obtained a bachelor's degree from the University of Tikrit/College of Computer Sciences and Mathematics Department of Computer Sciences in 2017. To complete their M.Sc. He moved to İstanbul/TÜRKİYE in 2017. He started his master's education at the Department of Information Technology at Altinbas University. To Complete their Ph. D. He moved to Karabük/TÜRKİYE in 2019. He started his Ph.D. education at the Department of Computer Engineering at Karabuk University.