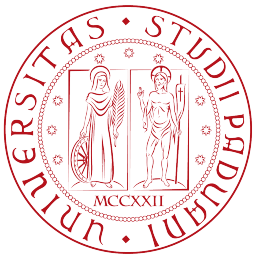

Università degli studi di Padova

Department of Information Engineering

Master Degree in Computer Engineering

Self-supervised Deepfake detection with Local Feature Exploration

Supervisor

Prof. Loris Nanni

Supervisor

Prof. Vitomir Štruc

Graduating Student

Elahe Sadat Soltandoost

Matricola: 2044004

Academic Year 2023/2024

Graduation Date 16/10/2024

"The best way to predict the future is to create it."

Abraham Lincoln

Contents

List of Figures

List of Tables

Il rilevamento dei deepfake rimane un compito difficile e critico nella sicurezza. Nessun modello eccelle in tutti i tipi di volti manipolati. Questa ricerca mira a scoprire l'importanza delle diverse parti del volto nel compito di rilevamento dei deepfake utilizzando Self-Blended Images (SBI). L'approccio SBI prevede la generazione di una maschera da un volto, la sua manipolazione e la sua fusione per creare un'immagine falsa. Estendiamo questa tecnica generando maschere distinte per occhi, naso e bocca. Quindi applichiamo il metodo SBI per modello e addestriamo tre modelli su compiti più complessi. Nella nostra implementazione, estraiamo l'input dell'ultimo livello completamente connesso in EfficientNet-04 per diversi generatori di maschere. Quindi, definiamo tre analisi delle componenti principali (PCA) e perfezioniamo un perceptron multistrato (MLP) per ogni tipo. In questo approccio, la capacità del modello viene sfruttata per concentrarsi su regioni facciali specifiche, il che ha il potenziale per migliorare il potere discriminante del modello. I nostri risultati indicano che, mentre l'approccio SBI esteso non migliora universalmente le prestazioni in tutti i set di dati, mostra un notevole miglioramento per un set di dati. Ciò evidenzia il potenziale del nostro metodo in determinati contesti e suggerisce aree per ulteriori perfezionamenti e ottimizzazioni.

Abstract

Deepfake detection remains a challenging and critical task in security. No single model excels across all types of manipulated faces. This research aims to discover the importance of different parts of the face in the deepfake detection task using Self-Blended Images (SBI). The SBI approach involves generating a mask from a face, manipulating it, and blending it back to create a fake image. We extend this technique by generating distinct masks for the eyes, nose, and mouth. Then we apply the SBI method per model and train three models on more complex tasks. In our implementation, we extract the input of the last fully connected layer in the EfficientNet-04 for different mask generators. Next, we define three principal component analyses (PCAs) and fine-tune a multi-layer perceptron (MLP) for each type. In this approach, the model's ability is leveraged to focus on specific facial regions, which has the potential to enhance the discriminative power of the model. Our findings indicate that while the extended SBI approach does not universally improve performance across all datasets, it shows notable improvement for one dataset. This highlights the potential of our method in certain contexts and suggests areas for further refinement and optimization.

1

Introduction

Nowadays, deepfake technology has garnered significant attention. This technology enables the creation of synthetic media by substituting an individual's appearance in images or videos with that of another person. This technology, originally popularized by an anonymous user under the alias "deepfake," has been used for both malicious purposes—such as eroding public trust and manipulating opinions—and potentially beneficial applications, like enhancing educational methods and enabling virtual interactions for individuals with disabilities [9].

The threats posed by deepfakes, including misinformation and identity theft, underscore the need for advanced detection methods. Recent advancements in deepfake detection involve sophisticated techniques such as Xception, capsule networks, and EfficientNet, which outperform earlier models based on shallow convolutional neural networks (CNNs) and basic facial features [10, 11]. Despite these improvements, deepfakes continue to present significant challenges for both automated systems and human evaluators [12, 13]. This emphasizes the ongoing need for robust detection tools and methods.

Moreover, deepfake detection algorithms increasingly rely on explainable artificial intelligence (XAI) techniques. These methods aim to provide clarity on the decision-making process of detection models, enhancing their reliability and trustworthiness [14]. The FST-Matching Deepfake Detection Model, for instance, improves detection efficiency by assessing visual concepts using the Shapley value [15]. Additionally, techniques that utilize source and target feature encoders help in demonstrating the realism of an image and differentiating between genuine and manipulated media [16].

The complexity of machine learning models utilized in deepfake generation and detection has introduced new challenges, particularly in the context of comprehending and clarifying their operation, as the digital landscape continues to develop.

1.1 Motivation

Deepfake technology has become increasingly sophisticated, allowing for the creation of highly realistic manipulated media that can deceive even the most discerning viewers. The inherent complexity of the deep learning models used in deepfake generation and detection often results in a lack of transparency, making it challenging to understand how these models make their decisions. This opacity can lead to issues related to fairness, trust, and accountability, especially in sensitive applications where it is crucial to comprehend the rationale behind model predictions.

The challenge of interpreting high-dimensional feature representations in deepfake detection models is similar to that seen in facial recognition systems, where there is a need for more transparent and understandable insights. The work of [17] demonstrates that specific directions in the embedding space can correspond to meaningful facial attributes [18]. This foundational approach can be extended to deepfake detection, where uncovering these interpretable directions could help us understand the specific features that differentiate authentic content from manipulated media.

By making deepfake detection more explainable, we gain valuable information about how detection models identify deepfakes, which can significantly assist human examiners who may wish to scrutinize and validate automated decisions. This increased transparency not only enhances the detection process but also supports human oversight, fostering greater trust in automated systems.

Given these advancements, our research aims to bridge the gap between complex model embeddings and human-understandable insights. To achieve this objective, we design a local feature discovery procedure that steers the deepfake detection tasks toward spatially local facial features. In other words, we devise a method that looks at isolated facial features, such as the eyes, the nose, or the lower face, at the time and then analyzes these local regions to determine whether the given facial image (or frame) is a deepfake or not. We achieve this by first perturbing the local facial regions through various data degradation technique and then capture the variations of the local regions through Principal Component Analysis applied in the embedding space of a pretrained deepfake detector. With the PCA transform we seek to uncover and visualize prin-

cipal components that correspond to local and thus interpretable facial features when used for deepfake detection. By applying "Principal Component Analysis" (PCA) to the embedding space of face recognition models, we seek to uncover and visualize principal components that correspond to interpretable facial features. This approach help demystify the internal workings of facial recognition systems and enhance their overall transparency.

1.2 Goals

To address the challenge of interpretability in deepfake detectors, our research is guided by the following objectives:

1. **Develop an Interpretable Feature Extraction Framework:** We aim to enhance the transparency of face recognition models by applying distinct Principal Component Analysis (PCA) models to different parts of the face. Each PCA model is specifically tailored to decompose high-dimensional feature vectors for particular facial regions, such as the eyes, nose, mouth, and chin. This approach allows us to capture and interpret the unique contributions of each facial part to the model's overall recognition process. By revealing these interpretable components, we seek to provide clearer insights into how the model makes its predictions and to facilitate a better understanding of the underlying features influencing recognition outcomes [18].

2. **Enhance Interpretability Through Visualizations:** We develop visualizations that illustrate the contributions of different facial regions (e.g., eyes, nose, mouth) to the model's decisions. In our approach, separate Principal Component Analysis (PCA) models are applied to different facial regions. This allows us to break down the feature vectors for each region into interpretable components specific to that part of the face. By visualizing these components, we can clearly demonstrate how each facial region contributes to the recognition process. This method enhances the transparency of the model by making the influence of distinct facial regions more interpretable, helping users understand how variations in each region affect recognition outcomes.

3. **Increase Confidence in Model Predictions:** Our model is expected to increase the confidence in the model's predictions. By providing clear, interpretable insights into the model's decision-making process, users and stakeholders can better trust the outcomes, especially in sensitive applications where understanding the rationale behind predictions is as important as the predictions themselves.

1.3 Thesis Structure

This thesis is organized into several chapters, each focusing on different aspects of deepfake technology, its implications, and detection methodologies:

This thesis begins with Chapter 2, which provides a comprehensive examination of deepfake technology, including its evolution, advantages, and associated risks, and reviews both traditional and modern detection approaches, highlighting the importance of developing robust strategies. Chapter 3 delves into the datasets utilized in deepfake detection, offering an in-depth analysis of Celeb-DF, FFIW, DFDC, and FaceForensics++, focusing on their characteristics, synthesis methods, and visual fidelity. Chapter 4 introduces the concept of Self-Blended Images (SBI) and details the process of creating and utilizing these images to improve deepfake detection, covering the objectives, key components, and methodologies for training and evaluating the SBI model. In Chapter 5, the thesis shifts focus to facial image classification techniques, discussing data preparation, facial landmarks, data augmentation, and the use of EfficientNet-B4 and Principal Component Analysis (PCA) in the classification process. Chapter 6 presents the experiments and results of the developed detection and classification models, evaluating their performance and comparing the findings with existing approaches. Finally, Chapter 7 concludes the thesis by summarizing the study's findings, highlighting significant contributions, and suggesting potential directions for future research in deepfake detection.

2

Background

This chapter provides a comprehensive examination of the historical background and evolution of media manipulation. This thesis explores the origins and impacts of deepfake technology, examines its various applications, assesses the potential risks and challenges it presents in the digital realm, and outlines the structure and focus of the subsequent chapters.

Figure 2.1 provides a visual representation of the fundamental phases required in generating deepfakes, aiding in the comprehension of the process.

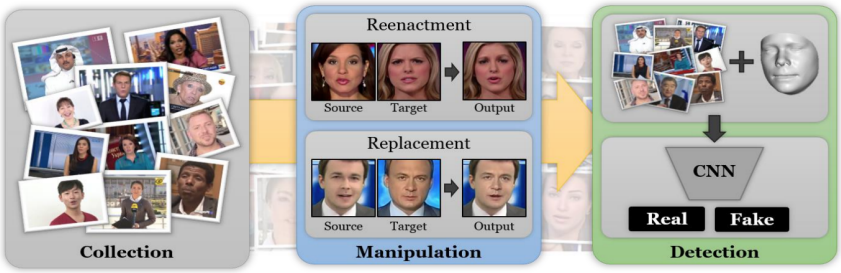

Figure 2.1: General process of generating deepfakes. Image Source[1].

Prior to the digital era, there has been a long-standing interest in manipulating media, including photos, audio, and video information [19]. Although modifying photographs has become reasonably easy with tools such as Adobe Photoshop, changing audio and video has traditionally posed more significant difficulties. During the initial era of cinema, the process of incorporating special effects into videos required painstaking frame-by-frame editing, which exemplified the laborious nature of the task [19].

The ubiquitous presence of personal computers has facilitated the inclusion of nearly all individuals in video editing, resulting in a significant rise in accessibility and democratization.

This accesibility has made it extremely easy to generate synthetic images and video today or manipulate existing imagery. Moreover, generating deepfake images or videos

has never been easier, allowing individuals with relatively basic computer knowledge to produce altered/manipulated and potentially harmful image content.

2.1 Benefits and Advancements

Deepfake technology offers numerous advantages for enterprises. AI-driven technology is currently being employed to transform a wide range of industries, including marketing, education, and entertainment. Below are a few beneficial applications of deepfake technology:

- **Low-Cost Video Campaigns**. With deepfake technology, marketers can create video campaigns without needing an in-person actor. Instead, they can purchase an actor's identity license and use previous digital recordings of the actor to create a new video. This can save time and money and also allow for easy edits to be made without the need for reshooting.

- **Hyper-Personalization**. Deepfake technology allows brands to provide customers with more personalized messaging and experiences based on their preferences. For instance, a brand can alter a model's skin tone in their marketing to better suit a customer's ethnicity or skin color, thus increasing inclusivity and reaching a broader market with their campaigns.

- **Bringing the Deceased Back to Life**. Deepfake technology can also captivate viewers or clients by delivering tailored recommendations and offers to fulfill their requirements. Reuters has developed an artificial intelligence-generated deepfake individual who is responsible for delivering the sports news summary. Furthermore, fashion companies are employing virtual fitting rooms, where customers have the ability to utilize deepfake technology to superimpose their own faces onto virtual models, allowing them to visualize how garments would appear on themselves.

2.2 Challenges and Risks of Deepfake Technology

The growing ubiquity of deepfakes in the digital sphere raises a number of concerns, one of which is the idea that deepfake technology will permanently blur the line between what is considered 'authentic' and what is considered 'fabricated' [20].

A smart production by Jordan Peele in 2018 had former President Obama offering a warning about deepfakes. This was one of the most notable examples of deepfake technology receiving broad notice. The word "deepfakes" refers to a broad category of manipulated material, including audio, video, and photo content, that is produced through the use of deep learning techniques [19]. Significantly, advances in AI have made it possible to create entirely computer-generated photos that remarkably mimic real photographs, making it harder to distinguish between real and fake.

Deepfakes possess the capacity to disrupt political debate, disseminate misinformation and propaganda, and damage individuals' reputations. Furthermore, the malicious exploitation of deepfakes with the purpose of deception and extortion can pose a threat to security and confidentiality. These consequences extend beyond the domains of politics, society, and economics, posing a threat to the stability of the economy, the operation of democratic processes, and the overall stability of society.

Legislation, technological progress, and digital media literacy are all being used to limit the spread of deepfakes. Nevertheless, the detrimental impacts of deepfake technology are continuously advancing, requiring a persistent emphasis on meticulousness and ingenuity.

2.3 Types of Deepfake

Deepfake technologies are categorized into four primary types: entire face synthesis, reenactment (including facial expressions and body motions), facial attribute manipulation, and face-swapping. Each type represents a distinct approach to generating or altering digital faces, with specific methods and challenges associated with each**??**.

2.3.1 Entire Face Synthesis

This approach generates synthetic faces that do not correspond to real individuals. It relies on learning latent representations from face datasets to create hyperrealistic personas. Early advancements include Radford et al.'s DCGANs, which introduced a stable generative model architecture without pooling layers, using batch normalization for improved image synthesis [21]. ProGANs by NVIDIA further enhanced quality through progressive image refinement from low to high resolution [22]. StyleGAN, an evolution of ProGAN, introduced adaptive instance normalization (AdaIN) to control style and pose in generated faces [23], and StyleGAN2 improved image quality and removed artifacts [24]. These models are widely used for creating synthetic face databases and avatars.

2.3.2 Reenactment

Reenactment involves transferring facial expressions or body motions from one person to another. Traditional methods used computer graphics, such as Blanz et al.'s 3D morphable models [25] and Thies et al.'s real-time facial expression transfer with RGB-D sensors [26]. Advances in neural networks include:

- Neural Textures

  This method integrates traditional graphics with learnable neural textures to produce photorealistic outputs. Early work by [27] used deferred neural rendering, while Fried et al. [28] focused on lip reenactment for altering dialogue in videos.

- Face2Face

  This method allows for the precise reproduction of face expressions. The procedures entail the procedure of correlating auditory attributes with specific oral configurations. Moreover, CycleGAN, along with its derivatives like StarGAN and RecycleGAN, are employed to enhance the transmission of facial emotions.These variations incorporate cycle-consistency and multi-domain translation to increase the process [29]. The ICface paper enhanced the control of pose and expression by utilizing conditional GANs. On the other hand, the Ordinal Ranking Adversarial Networks paper improved the precision of synthesis by integrating age and

expression intensity.

2.3.3 Facial Attribute Manipulation

This type alters specific facial characteristics, such as age, gender, or expression. Techniques include StarGAN. StarGAN is a domain-to-domain translation network that supports multi-domain training using mask vectors [30]. Xiao et al.'s ELEGANT model enhances multi-attribute translation with adversarial and reconstruction losses [31]. AttGAN focuses on high-quality facial attribute outputs with attribute classification constraints, while STGAN and URCA-GAN improve output quality and reduce artifacts [32]. BeautyGAN is noted for makeup style transfer with high realism [33], and SC-FEGAN translates sketches into hyperrealistic textures

2.3.4 Face-Swapping

Face-swapping involves exchanging faces between individuals while preserving facial expressions. Techniques include:

- Fast Face Swap

  Introduced by Korshunova et al., this method maintains head position and lighting conditions while transforming identity using a modified CNN with content and style loss functions [34]. RSGAN and FSNET employ latent spaces for face-swapping, with improvements in stability and resolution [35]. Sun et al. focused on identity obfuscation with parametric face autoencoders and GANs [36]. The recurrent neural network approach by [**?**] combines pose and expression from landmarks to generate face-swapped images with high temporal coherence.

- Face Replacement Using Internet Libraries

  The study of [37] the technique of face swapping, automates face replacement in photographs by selecting faces from a large internet library. The system creates seamless composite images by selecting candidates with similar appearance attributes to the input face. It can handle illumination and skin color variations. The method estimates lighting and average color within the replacement region,

aiming to generate realistic replacement results for different poses.

2.4 Deepfake Detection

Deep learning achieved great success in deepfake detection. Here we discuss some image and video deepfake detection models. Deepfake detection has achieved rapid success in recent years, thanks to deep learning. In this section, we categorize some of the most recent studies on this topic.

2.4.1 Traditional detection methods

Before learning methods, calculating the similarity of face or face patches betwen the target and the source image/ video was the main method for face swapping, bending, and face reenactment techniques. One of the most used approach for detecting sawpped face is used copy, move and splicing detection wich was used mostly between two images [38]. Bitouk et al. developed a method for identifying an input image and guaranteeing its privacy. Their method entailed querying a database to identify a visage that bore a striking resemblance to the face of the input source. Furthermore, the author worked to enhance the process of seamlessly integrating the identified face with the input face. This strategy is restricted by its inability to facilitate the seamless interchange of any two features.

The method developed by [39] utilizes facial characteristics such as eyes, teeth, and contours to identify deepfake and computer-generated faces. The technology identifies visual anomalies in computer-generated photos, accurately differentiating between genuine and altered faces. It improves the accuracy of deepfake identification by highlighting tiny inconsistencies that are often missed by humans.

2.4.2 Learning-based methods

The progress in deep learning has enabled the detection of Deepfake movies, employing computer vision techniques like as classification, image synthesis, and facial recognition. Over the past few years, numerous methods for detecting Deepfakes have been proposed,

which employ advanced technologies such as deep neural networks, convolutional neural networks, recurrent neural networks, and hybrid models.

Image detection models

Researchers have employed Convolutional Neural Networks (CNNs) to detect images, leveraging their advanced capabilities to autonomously acquire and extract hierarchical features from unprocessed photos, resulting in exceptional outcomes (Sengur et al., 2018). These models have played a crucial role in pushing forward the area of image recognition, providing exceptional performance in comparison to conventional approaches. The adaptability and efficacy of Convolutional Neural Networks (CNNs) in several applications, such as object identification, face detection, and the more recent deepfake detection, highlight their success[40]. Gowda and Thillaiarasu utilized modified CNN models, specifically ResNet and Xception, to identify counterfeit photos in a research study. They achieved performance of ResNeXt and Xception of 80% and 78%, respectively, but the best results were with the ensemble of ResNeXt and Xception of 93% [41].

This approach involved comparing the locations of the head and different face regions to discover inconsistencies. An further research paper proposed a methodology that utilizes neural networks to detect fraudulent GNA films. This approach entails utilizing preprocessing techniques to analyze the statistical properties of the images and determine their authenticity [42]. The researchers did a comparative analysis of the results and observed improvements in the recognition of fraudulent faces in images produced by people.

Later, researchers addressed the importance of models' ability to generalize. [43] for instance, discuss how models can generalize and perform well on datasets other than their training datasets. They used a non-public image dataset that contains 53,000 images extracted from YouTube videos with manipulated faces. Despite the high accuracy of ImageNet-trained models like AlexNet, VGG19, and Inception, they encountered challenges in distinguishing between fake and real faces in the new artifact dataset.

In addition to these strategies, research benefits from a hybrid approach that combines CNNs as feature extractors and classical machine learning tools for effectively predicting

deep fakes in images and videos [44] [45]. The suggested strategy in [45] makes use of a modified network topology and a two-stream approach. Based on GoogleNet[46], the face classification stream trains the model using both manipulated and real photos. Using a steganalysis feature extractor, the patch triplet stream analyzes features and records local noise residuals and low-level camera properties. The model also includes a Siamese network design and an enhanced DenseNet backbone network. Using paired learning, this detector successfully identifies false pictures produced by GANs that were not used in the training phase. The outcomes of the experiments show that this method is accurate in differentiating between authentic and counterfeit photographs[45].

Another innovative approach to deepfake detection is the use of Self-Blended Images (SBI), which enhances the training of deep learning models by generating synthetic samples that replicate common forgery patterns. A technique called Self-Blended Images (SBI) is used to detect deepfake videos by creating artificial fake samples that exhibit typical signs of fraud. These samples are utilized for the purpose of training deep learning models, therefore augmenting their capacity to identify faked media. The approach of the SBI has three distinct stages: Source-Target Generation, Mask Generation, and Blending. The Source-Target Generator (STG) generates novel data by combining sections of an image with itself, whereas the Mask Generator (MG) generates blending masks that combine images from both the source and destination sources. The training and evaluation approach for the deepfake detection model based on SBI is holistic, encompassing preprocessing, source-target augmentation, training, model validation, and inference technique. Using the SBI approach, deep learning models are trained more effectively, resulting in improved accuracy and resilience in detecting corrupted media. This paper also examines future research and its consequences for the field[7].

Video detection models

In recent years, deep learning techniques have demonstrated remarkable success in identifying deep fakes. Video compression, however, results in the loss of some frame information, prompting the development of alternative methods that operate directly with videos. In this section, we present several pertinent studies.

The authors employ an integrated approach to identify deepfake content in videos,

utilizing CNN models, feature selection techniques, and machine learning algorithms. Their model receives each video frame for the purpose of feature extraction and classification. The approach integrates picture characteristics from three Convolutional Neural Network (CNN) models into a feature vector. This vector is subsequently employed for feature selection and dimensionality reduction using Principal Component Analysis (PCA). A Support Vector Machine (SVM) categorizes the frame as either genuine or counterfeit. This approach attained a remarkable accuracy of 96.50% on the DFDC dataset, surpassing the performance of baseline end-to-end CNN models.

However, the approach used a shortened version of the dataset, potentially excluding the full range of variables present in the original dataset[44]. The authors of [43] integrate the super-resolution algorithm with deep learning techniques to classify counterfeit videos. This method utilizes image processing techniques to detect the disparities across various face regions and the orientation of the head. This approach has greatly enhanced the precision.

In addition, Li (2018) utilized novel methodologies to minimize the necessity of include negative training instances in the model. Using a deep learning approach, rather than identifying unique characteristics of fake and real videos, the researchers analyze the presence of face-wrapping artifacts in the videos to determine their authenticity. They demonstrated successful outcomes for two distinct deepfake datasets.

[47] developed an advanced deep learning system that detects deepfakes by examining both the auditory and visual elements. The model utilizes a Siamese network architecture to extract speech and face data. It includes a triplet loss function to differentiate between authentic and fake videos. The suggested technique achieved a 96.6% accuracy on the DeepfakeTIMIT dataset and an 84.4% accuracy on the DFDC dataset. [48] introduced a method for detecting fraudulent face recordings by analyzing the frequency of eye blinking, a vital physiological trait. The technology combines a convolutional neural network (CNN) with a recursive neural network (RNN) to track and analyze eye movement and blinking. A binary classifier is used to determine the different conditions of the eye. When utilized on a particularly curated dataset that emphasizes eye-blinking, the method showcased its efficacy in detecting fraudulent photos. [49] devised a Generative Adversarial Network (GAN) model that detects deepfakes by analyzing physiological

data, such as heartbeats. The model utilizes detector networks and registration layers to extract precise facial regions and biological signals, leading to an impressive accuracy of 97.3% across multiple publicly available datasets. Another point of view addressed by many researchers is analyzing the temporal sequence between frames of a video to identify fake videos. For instance, in reference [8], a novel approach is presented that employs the CNN model to extract frame features in the early stages.

Afterwards, the output is passed through the LSTM layers to retrieve the chronological order in which face manipulation takes place. In the previous section, a softmax classifier is employed to distinguish between authentic and fraudulent movies (2.2). Their model underwent evaluation using a dataset including 600 movies obtained from various sources. The model exhibited robust performance in accurately identifying video deepfakes..

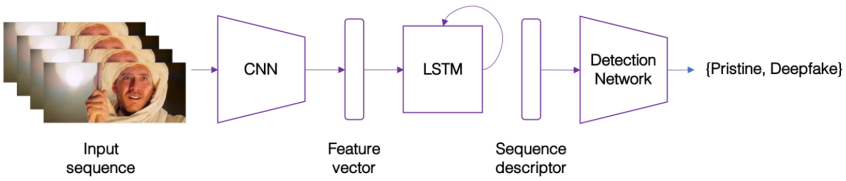

Figure 2.2: Convolutional neural network for spatial and temporal features analysis[2].

In another study, they replaced one face with another and benefit from color correction techniques such as Gaussian blurring, SURF descriptor . Finally to classify they use SVM. However, this method is unable to transfer facial expressions from the face to the sawed face[50].

3

Datasets

In this thesis, four distinct deepfake datasets are employed to rigorously evaluate the performance of deepfake detection models. These datasets are integral to the training and testing phases, each offering attributes that contribute to a comprehensive analysis of the detection algorithms' effectiveness.

3.1 Celebrities DeepFake dataset

The Celebrities DeepFake dataset, also known as the Celeb-DF dataset, is a collection of altered recordings that feature well-known individuals. This dataset has been employed in research concerning the societal implications of deepfake technology. The emergence of deepfake material, particularly in relation to personalities, presents novel ethical dilemmas and issues regarding the dissemination of fraudulent information and the infringement of privacy.

The Celeb-DF dataset is a notable progress in the field of DeepFake detection datasets. Celeb-DF improves upon earlier datasets by addressing concerns such as limited diversity and low-resolution movies. It establishes a higher standard by providing a wide variety of high-quality synthetic films. Celeb-DF is a carefully selected compilation of videos, which includes 590 authentic videos and 5,639 DeepFake recordings, amounting to more than two million frames. The 59 celebrities included in Celeb-DF were carefully selected to provide a diverse representation of genders, ages, and races, thereby providing a thorough coverage of diversity in the dataset.

The real videos in Celeb-DF are sourced from publicly available interviews on YouTube, ensuring a varied and realistic representation of facial expressions, orientations, lighting conditions, and backgrounds. The average length of all videos is approximately 13

seconds, with a standard frame rate of 30 frames per second. Regarding demographic distribution, 56.8% of the subjects in the real videos are male, and 43.2% are female, with further breakdowns across different age groups and ethnic backgrounds for a detailed representation. The age distribution is as follows: 8.5% of the population is aged 60 and above, 30.5% is between the ages of 50 and 60, 26.% is in their 40s, 28.0% is in their 30s, and 6.4% is younger than 30. The dataset comprises 5.1% Asians, 6.4% African Americans, and 88.1% Caucasians, demonstrating a conscious endeavor to incorporate a diverse array of ethnicities and foster inclusiveness within the dataset.

Synthesis Methodology

The DeepFake videos in Celeb-DF are generated using an advanced synthesis algorithm that addresses specific visual artifacts observed in existing datasets[3]. Initially, the basic DeepFake synthesis algorithm typically produced low-resolution faces (e.g., 64 × 64 or 128 × 128 pixels). In Celeb-DF, this has been improved by enhancing the resolution of the synthesized faces to 256 × 256 pixels. This enhancement is achieved through encoder-decoder models with increased layers and dimensions, balancing training time with improved synthesis quality Figure 3.1.

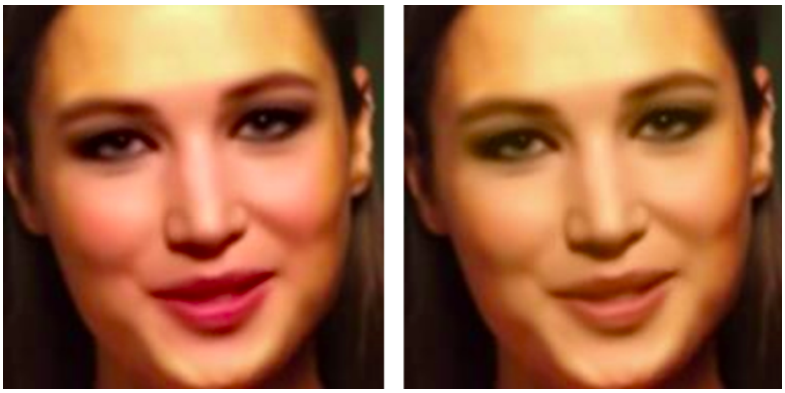

Figure 3.1: The visual sample of the Synthesis images, with and without corel correction from right to left.Image source [3]

Color mismatch between the synthesized donor's face and the original target's face is significantly reduced in Celeb-DF through rigorous data augmentation and post-processing techniques. During each training epoch, colors of training faces are randomly perturbed to force the neural networks to synthesize images with accurate color patterns. Additionally, a color transfer algorithm is applied between the synthesized donor face

and the input target face to minimize discrepancies.

Improvements in mask generation are another critical aspect of Celeb-DF. Unlike previous datasets where face masks were either rectangular or based on convex hulls of landmarks, Celeb-DF employs a refined approach. It synthesizes a face with more surrounding context to ensure complete coverage of the original facial parts after warping. A smoothness mask is then generated based on landmarks and interpolated points, enhancing the integration of the synthesized face with the target video frame (see Figure 3.2).

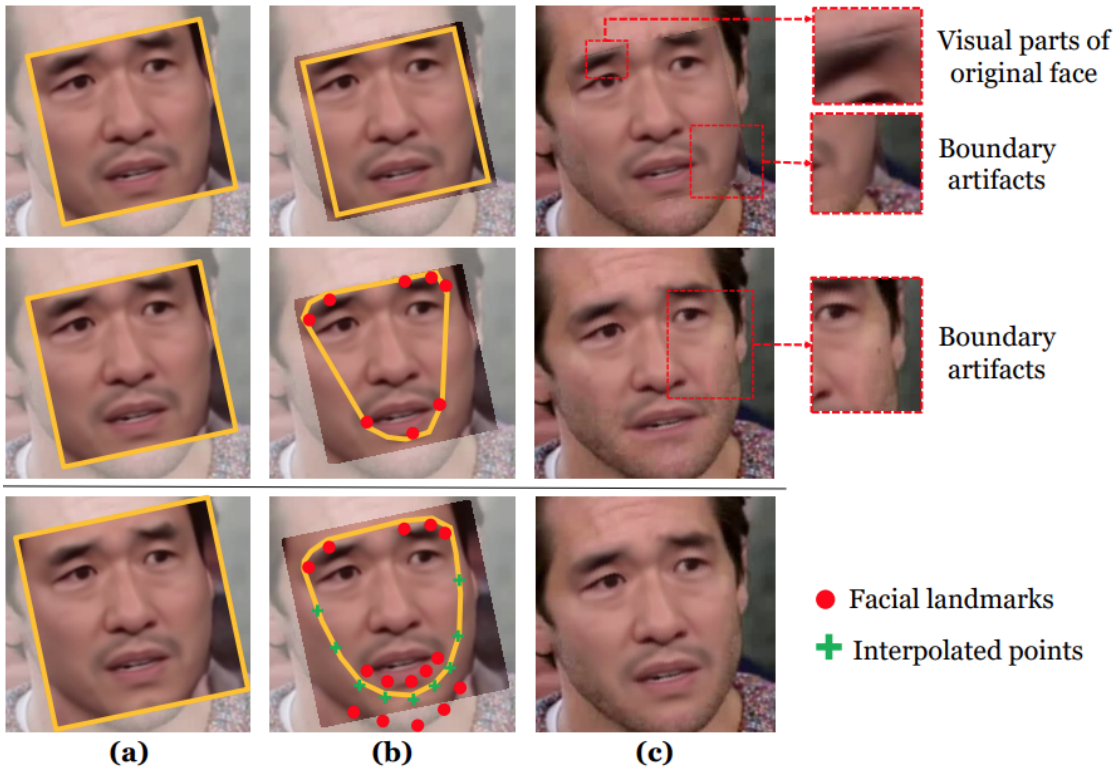

Figure 3.2: Mask Generation in Datasets and Celeb-DF (a) Initial alignment of the synthesized face with the target's face. (b) Smooth and accurate mask creation around facial features. (c) Final synthesis result integrating synthesized face into the original video frame. Image source [3].

Temporal flickering, a common issue in DeepFake videos, is mitigated in Celeb-DF by incorporating temporal correlations among detected face landmarks. A Kalman smoothing algorithm filters the temporal sequence of face landmarks, reducing imprecise variations across frames and ensuring smoother transitions between facial expressions.

3.1.1 Visual Quality Evaluation

The refinements in synthesis methodology for Celeb-DF lead to improved visual quality of DeepFake videos, as confirmed by quantitative evaluation metrics. The Mask-SSIM (Masked Structural Similarity Index Metric) score, which measures the similarity between the head regions (including face and hair) of DeepFake and original video frames, demonstrates superior performance compared to previous DeepFake datasets. Celeb-DF achieves higher Mask-SSIM scores, indicating better image quality and reduced identity changes from the target to the donor[3].

3.2 Face Forensics in the Wild 10K

The FFIW10K dataset, short for Face Forensics in the Wild 10K, represents a significant advancement in the field of deepfake detection datasets[51]. This dataset is specifically designed to overcome the limitations of existing benchmarks by providing a more comprehensive and realistic benchmark for evaluating deepfake detection algorithms. FFIW10K consists of a curated collection of both synthetic and real videos, totaling 20,000 clips and spanning over 7.2 million frames. The dataset is meticulously constructed to capture diverse and challenging scenarios of face manipulation in real-world settings.

To gather pristine videos for manipulation, the dataset creators sourced a large collection of high-resolution videos from YouTube. These videos were selected using diverse keyword queries in multiple languages to ensure a broad representation of human faces and scenarios. From the initial pool of 4,000 raw videos, each video was segmented into uniform clips, and a random 12-second sequence was selected to avoid bias. The selection process filtered out static or crowded scenes and those with minimal human facial presence, resulting in approximately 12,000 sequences suitable for subsequent manipulation. Some visual examples of the dataset are shown in Figure 3.3.

To ensure the diversity and quality of the manipulated videos, FFIW10K employs an automated quality assessment network. This network evaluates the visual fidelity of each manipulated face and discards low-quality synthetic faces. Only videos with high-quality synthetic faces are retained, resulting in a final dataset of 10,000 high-fidelity synthetic

videos. This rigorous curation process not only enhances the dataset's realism but also reduces human bias and ensures scalability.

FFIW10K is annotated comprehensively at both face and video levels, providing detailed labels for over 3.2 million real faces and 1.1 million synthetic faces across 3,600 individuals. These annotations enable precise evaluation and benchmarking of deepfake detection algorithms, facilitating robust analysis under real-world conditions. Moreover, the dataset is split into distinct training, validation, and test sets to support fair evaluation and comparison of detection models.

Figure 3.3: Visual examples of FFIW dataset. Image source [4]

3.3 DeepFake Detection Challenge

The DeepFake Detection Challenge (DFDC) Dataset is a pivotal resource developed to advance the detection of manipulated video content, specifically deepfakes. Created by Facebook (now Meta) and Kaggle as part of the DeepFake Detection Challenge, the dataset encompasses over 100,000 video clips. These videos feature a broad array of deepfake manipulation techniques, including face swapping and facial reenactment, alongside authentic footage. This extensive collection provides a diverse sample of individuals across various ages, ethnicities, and contexts, making it an invaluable asset for developing and evaluating deepfake detection algorithms. Some visual example of this dataset is shown in Figure 3.5

The dataset is meticulously annotated with labels indicating whether each video is "real" or "fake." It includes both high-quality and low-quality deepfakes, reflecting the wide range of video manipulation techniques and their varying degrees of sophistication. This diversity introduces real-world challenges for detection models, ensuring they are robust and effective across different types of manipulations and video qualities.

Ethical considerations were central to the creation of the DFDC dataset. The videos

were generated with explicit consent from participants, and no unauthorized or private material was used. The focus was on synthetic content to maintain ethical standards while providing a comprehensive resource for research and development.

Figure 3.4: Visual example of DFDC dataset. Image source [5].

The DFDC dataset not only serves as a benchmark for evaluating deepfake detection technologies but also plays a critical role in pushing forward the boundaries of media verification. It is widely utilized in academic research, industry applications, and competitive challenges to enhance the accuracy and reliability of algorithms designed to identify manipulated media. Available through Kaggle, this dataset is a cornerstone for those working to combat digital misinformation and enhance media authenticity.

3.4 FF++ (FaceForensics++)

A fundamental contribution to the field of deepfake detection is the FaceForensics++ dataset, an extension of the original FaceForensics dataset [5]. Some visual samples of the real and fake images are shown in Figure **??**. This new iteration, named FF++, significantly enhances the scale and diversity of manipulated facial images available for research, thereby enabling more robust and comprehensive evaluations of forgery detection algorithms. To simulate realistic scenarios of facial manipulation, FF++ incorporates four distinct subdatasets, each employing advanced manipulation techniques (Figure 3.6).

Subdatasets

- **DeepFakes**: The term "Deepfakes" has become synonymous with face replacement through deep learning methods. In FF++, the Deepfakes subdataset leverages

autoencoders trained on large collections of source and target face images. These autoencoders encode and decode facial features to seamlessly replace faces in target videos. The implementation utilizes a modified version of the faceswap github framework, which automates data loading and model training to efficiently generate manipulated videos.

- **Face2Face**: Face2Face is a facial reenactment system that transfers expressions from a source video onto a target video while preserving the target person's identity. This approach involves generating dense facial reconstructions from input video streams and applying these reconstructions to re-synthesize facial expressions under varying lighting and expression conditions. In FF++, Face2Face adaptations include fully automated keyframe selection and expression tracking, enhancing the scalability and realism of generated manipulations.

- **FaceSwap**: FaceSwap is a graphics-based approach that extracts facial regions from a source video and integrates them into a target video using blendshape models and texture mapping techniques. The method involves detecting facial landmarks, fitting 3D templates, and blending rendered models with target images to achieve realistic face swapping effects. FF++ incorporates FaceSwap with enhancements to handle varying facial poses and expressions, ensuring robust manipulation outputs.

Figure 3.5: Visual example of FF++ dataset. Image source[5].

- **NeuralTextures**: NeuralTextures introduces a rendering approach that learns neural textures from original video data to realistically manipulate facial expressions. This technique uses photometric and adversarial losses during training to refine facial textures and expressions. In FF++, NeuralTextures integrates with

Face2Face's tracking module to modify specific facial regions while preserving overall facial geometry, enhancing the fidelity of manipulated videos.

Figure 3.6: This figure illustrates two key processes in facial image editing: face swapping and facial reenactment. Face swapping involves altering identity by replacing one face with another, a technique now widely accessible on mobile devices. Facial reenactment modifies expressions by transferring the facial expressions from one person to another. Image source [6].

Annotation and Utilization

FF++ is extensively annotated with both face-level and video-level labels, delineating manipulated and pristine segments. This annotation scheme facilitates precise evaluation of forgery detection models and enables comparative studies against other benchmark datasets. The dataset is partitioned into training, validation, and test sets, ensuring unbiased evaluation and robust performance metrics across various detection methodologies.

4

Self-Blended Images (SBI) for Deepfake Detection

4.1 Introduction

One promising approach to deepfake detection is the use of Self-Blended Images (SBIs)[7]. The SBI technique introduces a novel strategy for generating synthetic fake samples that contain common forgery traces. These synthetic samples are then used to train deep learning models, in a discriminative manner enhancing their ability to detect a wide range of manipulated media. This chapter provides an in-depth of the SBI method, its underlying principles, and its impact on improving the accuracy and robustness of deepfake detection.

4.2 Overview of the SBI Model

The detection of manipulated images and videos has traditionally relied on supervised learning methods, where models are trained on labeled datasets containing both authentic and fake samples. However, these methods face significant challenges due to the ever-evolving nature of deepfake techniques and the scarcity of high-quality labeled data. The need for more generalized and robust detection mechanisms has led to the exploration of alternative approaches, such as self-supervised learning and synthetic data generation.

Self-supervised learning (SSL) is a subset of unsupervised learning where the data itself provides the supervision needed for training. It involves automatically generating labels from the data based on certain inherent properties or structures within it. This method allows models to learn useful representations from unlabeled data, which can later be

fine-tuned for specific tasks with minimal labeled data. SSL has shown significant success in fields such as computer vision and natural language processing, enabling models to leverage vast amounts of available data without the need for extensive manual labeling [52, 53, 54, 55].

The SBI method addresses these challenges by creating self-blended images that simulate common forgery traces. By blending parts of an image with itself, the SBI technique generates a variety of synthetic samples that exhibit characteristics of manipulated media. These samples are then used to train classifiers, encouraging them to learn more generic representations that are effective across different types of forgeries. As shown in Figure 4.1, the overall structure of the SBI model consists of three main components: the Source-Target Generator (STG), the Mask Generator (MG), and the Blending step.

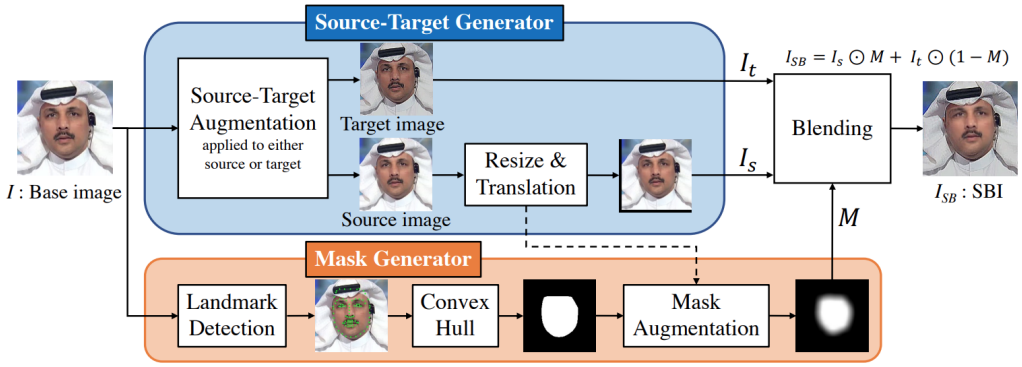

Figure 4.1: Overview of the SBI Model Architecture. Image source [7].

4.2.1 Objectives

The core idea behind SBI is to leverage self-supervised learning, where the model generates its own training data by blending parts of images within the same image. This process introduces realistic forgery traces, allowing the model to learn and identify subtle manipulations that are commonly found in deepfakes. By training the model with these synthetic samples, SBI aims to develop a more generalized understanding of forgery patterns, making it more effective in detecting manipulated content across diverse datasets.

4.3 Methodology

The Self-Blended Images (SBI) technique involves creating synthetic data by blending parts of an image with itself to produce fake samples that mimic typical forgery patterns. These synthetic samples are then used to train detection models, enhancing their ability to recognize deepfake manipulations. The SBI methodology comprises three key steps: Source-Target Generation, Mask Generation, and Blending. Each step plays a crucial role in creating synthetic samples with diverse characteristics for training deepfake detection models.

The Source-Target Generator (STG) is a crucial element of the SBI approach. The process starts by replicating the input image I to generate faux source and target images. The STG utilizes random alterations to introduce statistical discrepancies among these images. The changes encompass several operations such as moving RGB channel values, affecting hue, saturation, brightness, and contrast, and altering image frequency through downsampling or sharpening techniques. These modifications guarantee that the artificial images are diverse and authentic, replicating many forms of counterfeit patterns.

In addition, the STG adjusts the size of the source image by randomly selecting height H_r and width W_r values from a uniform distribution $U[umin, umax]$. Here, $u_{\min}$ and $u_{\max}$ represent the minimum and maximum scaling factors, respectively, used to resize the image. The procedure of resizing is crucial in order to preserve the alignment with the original dimensions of the image. This may be achieved by either adding zero-padding or cropping the image from the center to match the precise size of the original image. Ensuring the structural integrity of the image is maintained while adding variations is of utmost importance in this stage.

In order to increase the variety and intricacy of the artificial data, the STG incorporates translation. This is accomplished by establishing a translation vector $t = [t_h, t_w]$, where t_h and t_w are computed using the dimensions H and W of the original image. These modifications together generate artificial images with different attributes, enhancing the resilience of deepfake detection models by offering a more extensive range of training instances.

4.3.1 Mask Generator (MG)

The Mask Generator (MG) component in the SBI technique is essential for creating blending masks that merge source and destination images. At first, the MG uses a landmark detector to anticipate face areas and starts a mask by calculating the convex hull using these landmarks [56]. Afterwards, the mask is subjected to deformations by landmark transformations, which are comparable to the ones employed in BiGAN. In order to increase the variety of masks, the MG utilizes elastic deformations and uses two Gaussian filters with different settings to achieve smoothness.

After the first smoothing process, all pixel values that are less than 1 are changed to 0. This enables erosion or dilation to be performed dependent on the sizes of the filter kernels. This procedure guarantees that the masks possess diverse and authentic bounds, which is crucial for generating credible synthetic images. Ultimately, the MG adjusts the blending ratio by multiplying the mask with a constant r that falls within the range of (0, 1]. The value of r is randomly selected from the set of numbers {0.25, 0.5, 0.75, 1, 1, 1}. The difference in blending ratios increases the diversity of the synthetic samples, hence improving their effectiveness for training detection algorithms.

4.3.2 Blending

The last stage of the SBI approach entails merging the source image I_s and the target image I_t using the blending mask M. The self-blended image I_{SB} is calculated by combining the source image I_s and the target image I_t using a blending factor M. The blending operation is performed by multiplying the source image by M and the target image by $(1 - M)$, and then adding the two results together. The blending procedure guarantees that the synthetic images accurately replicate the attributes of genuine forgery patterns. The blending method creates self-blended images by merging various parts of the same image. This process efficiently imitates common counterfeit patterns, which helps improve the training of detection models.

4.4 Key Components

The SBI approach has many essential elements aimed at improving the identification of altered images and videos:

- **Image Blending for Synthetic Sample Generation:**The Image Blending Algorithm produces self-blended images by merging several areas inside a single image. The generateed synthetic samples imitates common counterfeit patterns by merging portions of a image with itself. This procedure is vital for producing a wide range of authentic training data, which is necessary for training resilient detection models.

- **Analysis of Forgery Traces**: This component detects and integrates typical signs of fraud into the artificial samples. The traces mentioned encompass several abnormalities and anomalies commonly detected in altered, such as irregular lighting, artificial shadows, and misaligned edges By integrating these subtle indications, the artificial samples become more authentic and demanding, enhancing the model's capacity to identify counterfeit items.

- **Training Pipeline**: The artificial samples produced by the SBI process are utilized to train machine learning models. This training process is designed to enhance the models' capacity to differentiate between genuine and altered material, hence boosting overall accuracy in detection. The pipeline incorporates diverse data augmentation approaches to include supplementary heterogeneity in the training data, hence enhancing the resilience of the models.

4.5 Training and Evaluation Methodology

The training and evaluation methodology for the SBI-based deepfake detection model is comprehensive, involving several key stages: preprocessing, source-target augmentation, training, model validation, and inference strategy.

4.5.1 Preprocessing

To process the input data, Dlib [56] and RetinaFace [57] are employed to extract facial landmarks and bounding boxes from each video frame, respectively. Specifically, the Dlib face landmark shape predictor [58] with 81 points is used. During the training phase, the facial region is cropped using a margin that is randomly determined and can range from 4% to 20%. However, during the process of drawing conclusions or making deductions, a consistent margin of 12.5% is used. It should be noted that face landmarks are not required throughout the inference process. Thus, just the RetinaFace algorithm[57] is employed at that juncture. The preprocessing procedure ensures that the facial regions are evenly aligned and normalized, which is crucial for training accurate detection models.

4.5.2 Training

The training approach utilizes the cutting-edge convolutional network architecture EfficientNet-b4(EFNB4) [59], which has been pretrained on ImageNet[60]. This architecture has demonstrated exceptional performance in several computer vision applications, including the identification of deepfake images. The model undergoes fine-tuning using the Stochastic Weight Averaging (SWA) optimizer [61] for 100 epochs. The batch size is set to 32 and the starting learning rate is 0.001. It should be noted that a Sigmoid activation function is utilized on the last layer in order to standardize the output within the range of 0 to 1.

4.5.3 Blending Masks

Building upon the Self-Blended images (SBI) approach, they apply a blending procedure to combine the modified masks with the original images. By creating many masks and using Gaussian and other modifications, they seamlessly merge these masks with the original facial images. This procedure entails combining the masks with the face areas through the utilization of alpha blending and other image fusion techniques in order to generate modified iterations of the original images.

The blending method guarantees the smooth integration of the altered masks, which individually represent distinct forms and configurations, with the original face characteristics. This stage enables us to assess the influence of these masks on the analysis and categorization of face features. The images presented in Figure 4.2 exemplify instances of modified images created by blending. These images serve to demonstrate the efficacy of our technique in modifying facial characteristics while maintaining the overall coherence of the image.

Figure 4.2: Examples of facial images after blending with manipulated masks. The images illustrate how different masks, applied to specific facial regions, are integrated with the original features using blending techniques. The blending process showcases the impact of various mask shapes and configurations on the final image, providing insights into the effects of mask manipulation on facial feature analysis. Image source [7].

4.5.4 Inference Strategy

During the process of inference, predictions are generated at the individual frame level and then combined through averaging to create predictions at the video level. This methodology guarantees the coherence of the model's forecasts over many frames within a given video, hence enhancing the overall precision and dependability of deepfake identification.

4.6 Conclusion

This chapter has presented an overview of the Self-Blended Images (SBI) technique for deepfake detection. By generating synthetic samples that simulate common forgery

traces, SBI enhances the training of deep learning models, improving their accuracy and robustness in detecting manipulated media. The methodology, key components, training, and evaluation strategies have been discussed in detail, highlighting the effectiveness of the SBI approach across various benchmark datasets. Future research and implications for the field have also been considered, underscoring the significance of SBI in advancing deepfake detection technologies.

5

Methodology for Facial Image Classification

5.1 Introduction

Our approach to deepfake detection in this chapter expands upon the Self-Blended Images (SBI) model from the previous chapter. Our main objective is to enhance the interpretability of the detection process by focusing on local picture regions. A novel approach called Local Feature Analysis SBI (LFA-SBI) is proposed to discover feature space directions in the embedding space of a deepfake detection model trained using the original SBI technique. The given directions are indicative of deepfake artifacts that are observed in facial features that are spatially localized, such as the eyes, nose, or mouth.

The fundamental concept of LFA-SBI is to manipulate various forms of local image disturbances inside certain face areas and subsequently represent these changes in the embedding space of a pre-trained deepfake detector using Principal Component Analysis (PCA). Given that PCA effectively captures the directions with the greatest variance in the data, we exploit this characteristic to determine the embedding space directions that precisely correspond to the changes in the desired local picture regions. This method of self-supervision allows the model to precisely identify the specific local features that contribute to the categorization decision.

This work aims to examine the effects of utilizing various facial masks on the extraction and categorization of facial characteristics using Principal Component Analysis (PCA) and Multi-Layer Perceptrons (MLPs). This study aims to evaluate the impact of masking several facial regions, namely the eyes, nose, and mouth, on the efficacy of facial feature analysis and classification. This methodology has the dual objective of accurately categorizing facial characteristics and identifying the specific areas of the

face that have been modified in artificial or manipulated photographs. This study contributes to Explainable AI (XAI) by elucidating the precise facial features that impact classification decisions.

After identifying the directions in the feature space, we proceed to train individual Multilayer Perceptrons (MLPs) for each local region. This enables the model to independently determine whether the input face is genuine or modified by self-blending. In this procedure, predictions are generated for the input image based on each trained Multilayer Perceptron (MLP), hence enhancing the transparency and interpretability of the decision process in relation to each studied facial region.

The technique, as depicted in Figure 5.1, comprises the subsequent essential phases:

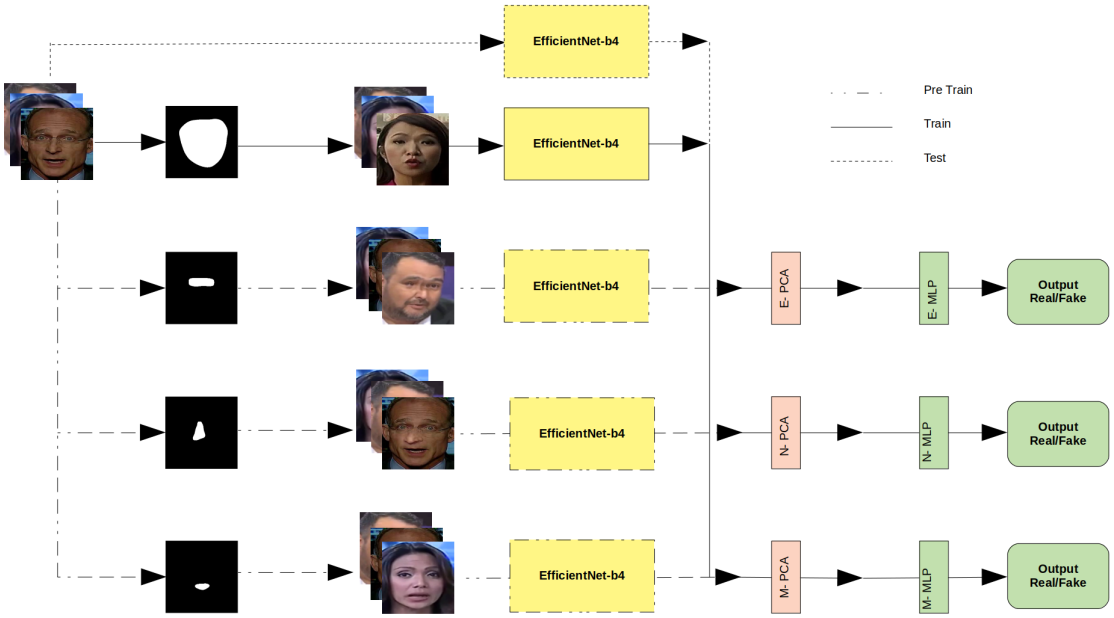

Figure 5.1: The main structure of the model. In the first step, the face part masks are extracted, then applied to the dataset to generate related PCA in the pre-train phase. Next, in the training phase, the whole face masks are used, and after passing through the EfficientNet-b4, PCAs are applied to the output of EfficientNet-b4 before the last fully connected layer. Using the output of each PCA, an MLP is trained. For the test phase, the faces are directly fed into the network.

1. **Face Parts Mask Generation**: Generating and using masks to divide facial images into discrete parts, including the eyes, nose, and mouth.

2. **Principal Component Analysis (PCA)**: The application of Principal Component Analysis (PCA) to the masked face parts serves to mitigate the influence

of noise by prioritizing the principal components that yield the highest variance, therefore improving the quality of the data.

3. **MLP Training**: Training MLP models on PCA-transformed data to assess the influence of various masks on classification performance.

4. **Evaluation and Explainability**: Evaluating the trained model to determine its effectiveness in classifying facial features using applied masks, and employing Explainable AI methods to detect the specific facial components that were modified in manipulated images, thus offering clarification on the modifications made to the face.

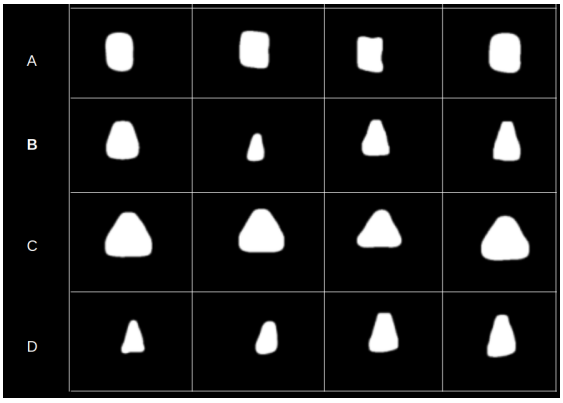

Figure 5.2: Masks for the nose region arranged in a 4x4 grid. Each row illustrates a different shape of the mask: (A) rectangular masks, (B) triangular masks, (C) wider triangular masks, and (D) original masks. The columns within each row represent variations in mask dimensions and configurations. This layout demonstrates the diversity in mask shapes applied to the extracted facial landmarks, showcasing the range of geometric configurations used in our analysis [8].

Through an analysis of the impacts of various masks, our objective is to construct a model that not only improves the accuracy of classification but also offers a comprehensive understanding of the particular changes in facial features detected in deepfake material. By enhancing the interpretability of the detection process and emphasizing the facial characteristics most pertinent to classification decisions, this adds to the broader objectives of XAI.

Our approach diverges from the conventional SBI methodology, which employs Dlib and RetinaFace to extract facial landmarks and bounding boxes from every video frame, by explicitly focusing on extracting independent facial components. This modified procedure integrates sophisticated data augmentation principles influenced by SBI, with

the goal of improving the resilience and effectiveness of the model. In addition to enhancing the accuracy of deepfake detection, the resulting LFA-SBI approach provides a clear understanding of how local facial characteristics contribute to the classification of manipulated photos.

5.2 Data Preparation

In contrast to the approach outlined in the SBI paper, which utilizes Dlib and RetinaFace for extracting facial landmarks and bounding boxes from each video frame, our methodology specifically targets the extraction of distinct facial parts. The adapted process is as follows, incorporating advanced data augmentation techniques inspired by SBI to enhance model robustness and performance:

5.2.1 Facial Landmarks and Bounding Boxes

- **Landmark Extraction**: We use Dlib, a C++ library with Python bindings, to extract facial landmarks from each video frame. Dlib is a modern C++ toolkit originally developed in C++ but also provides Python bindings for easy integration. It contains machine learning algorithms and tools for creating complex software in C++ to solve real-world problems. It includes tools for facial landmark detection, which utilizes a pre-trained model to identify key points on the face, such as the eyes, nose, mouth, and jawline. The facial landmark detector in Dlib uses an ensemble of regression trees to predict the location of facial landmarks. It begins by detecting the face in the image and then applies a shape predictor to identify the precise positions of 68 landmarks (5.3). These landmarks provide a detailed map of the facial structure, which can be used for various facial analysis tasks.

  To address these challenges, RetinaFace provides an advanced solution for bounding box extraction. RetinaFace employs a Feature Pyramid Network (FPN) to detect faces at multiple scales and an anchor-based approach to improve bounding box accuracy [57]. This method enhances detection performance by using dense predictions and integrating facial landmark localization to refine the bounding boxes further [62].

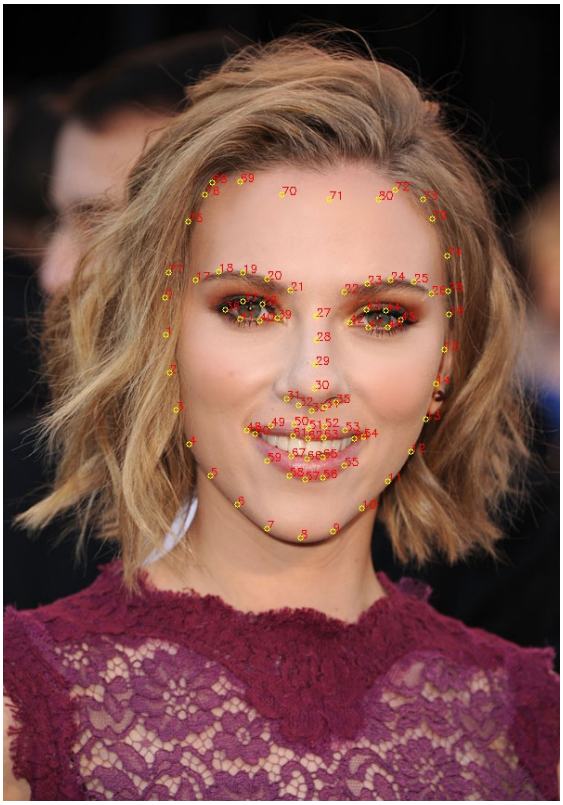

Figure 5.3: Visualizing the 68 facial landmark coordinates from the iBUG 300-W dataset [8].

From each video, 32 frames are extracted to ensure a comprehensive analysis. Unlike the SBI approach, which uses landmarks for the entire face, our process focuses on landmarks corresponding to specific facial parts: eyes, nose, and mouth.

- **Bounding Box Extraction**: Bounding box extraction is a critical step in image analysis, particularly for face detection tasks. It involves delineating a rectangular area around detected faces to isolate them from the rest of the image. This process ensures that subsequent analysis focuses on the regions of interest, improving the accuracy of tasks such as feature extraction or classification. Accurate bounding box extraction is essential for applications like facial recognition and deepfake detection, where precise localization of facial features is crucial. Techniques such as the Single Shot MultiBox Detector (SSD) [63] and Faster R-CNN [64] have been traditional methods for face detection, but they often face limitations in handling faces at varying scales and orientations.

  To address these challenges, RetinaFace provides an advanced solution for bounding box extraction. RetinaFace employs a Feature Pyramid Network (FPN) to

detect faces at multiple scales and an anchor-based approach to improve bounding box accuracy [57]. This method enhances detection performance by using dense predictions and integrating facial landmark localization to refine the bounding boxes further [62].

Similar to the SBI methodology, we use RetinaFace to obtain bounding boxes around the face, providing the initial region of interest for further analysis.

5.2.2 Facial Part Cropping

Instead of cropping the entire face, we create masks for specific facial parts using extracted landmarks. These masks isolate the nose from other facial features and are uniquely shaped to reflect the diversity of nose shapes and positions. Figure 5.2 illustrates various masks tailored for the nose, each associated with specific landmarks from different face images. As you see, there are 16 masks arranged in a 4x4 grid. Each row represents a different shape of the mask used for the nose, and each column shows variations within that shape. These masks have undergone Gaussian and other manipulations and are ready to be applied and blended onto source images to create altered versions. The varying geometric configurations of these masks demonstrate how different shapes can be applied to the landmarks. This approach enhances the flexibility of our analysis, capturing a wide range of variations in the nose region and providing more detailed insights into facial feature dynamics.

5.2.3 Data Augmentation

We implement data augmentation strategies to improve the robustness of our models. These techniques include introducing various degradations that generate artifacts in the images. These artifacts are designed to simulate potential distortions and alterations, enabling our detectors to learn and recognize such variations more effectively. The specific augmentation methods are as follows:

- **Color Transformations**:
  - **Random Shifts in RGB Channels**: Introduce color variability by adjusting the values of the red, green, and blue channels.

  - **Hue, Saturation, Value Adjustments**: Modify hue, saturation, and value to simulate different lighting conditions and color intensities.

- **Frequency Transformations**:
  - **Downsampling**: Reduce the resolution to simulate lower quality or zoomed-out views.
  - **Sharpening**: Enhance image sharpness to introduce high-frequency details.

- **Geometric Transformations**:
  - **Resizing**: Change the size of the source image to create blending boundaries and simulate different spatial relationships between facial features.
  - **Translations**: Shift the position of the resized image to reproduce alignment mismatches.

- **Blending**:
  - **Mask Application**: Apply the manipulated masks to the source images to create blended images, integrating the augmented source images to simulate realistic and challenging scenarios.

This approach allows us to generate a diverse set of training samples, improving the model's ability to handle various distortions and alterations in facial images.

5.2.4 Blending Masks

Inspired by the Self-Blended Images (SBI) technique, we implement a blending process to integrate the manipulated masks with the source images. These manipulations include random shifts in RGB channels, hue, saturation, value, brightness, and contrast to create color inconsistencies. Additionally, we apply frequency transformations such as downsampling or sharpening. The source image is resized with dimensions sampled from a uniform distribution, then zero-padded or center-cropped to the original size, followed by random translations.

This blending process ensures that the manipulated masks, each reflecting different shapes and configurations, seamlessly integrate with the original facial features. This

Figure 5.4: Examples of facial images after blending with manipulated masks. From left to right, the columns represent the mouth, nose, and eye regions, respectively. The top row shows the original images, while the bottom row displays the images after applying the manipulated masks. This visual comparison illustrates the impact of mask manipulation on each specific facial region, providing insights into the effects of different mask shapes and configurations on facial feature analysis.

step allows us to evaluate the impact of these masks on facial feature analysis and classification. The figures provided in Figure 5.4 showcase examples of manipulated images after blending, demonstrating the effectiveness of our approach in altering facial features while preserving overall image coherence.

To illustrate the blending process, Figure 5.5 shows the intermediate steps, including the original frame, the cropped face region, the applied mask, and the final blended result. This visualization offers a comprehensive view of how each step contributes to the creation of a manipulated image, highlighting the transformations and blending techniques used.

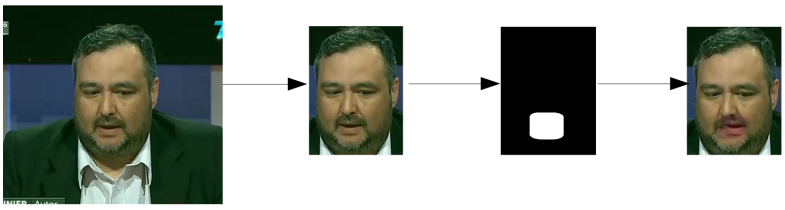

Figure 5.5: Visualization of the blending process. Each column represents a different facial region (mouth, nose, eyes). The rows depict the sequence from the original frame, the cropped face region, the applied mask, to the final blended result. This step-by-step visual breakdown demonstrates the process of mask application and blending, illustrating the transformations applied to create the final manipulated image.

5.3 EfficientNet-B4

EfficientNet-B4 is a model from the EfficientNet family, which represents a significant advancement in the field of deep learning and computer vision. EfficientNet models are designed to achieve high accuracy with fewer parameters and computational resources compared to traditional neural network architectures. EfficientNet is a family of models introduced by Google AI that are optimized for both accuracy and efficiency. The key innovation behind EfficientNet is a compound scaling method, which uniformly scales the depth, width, and resolution of the network to achieve better performance with fewer parameters. For this study, we utilize a pretrained version of EfficientNet-B4, which has been trained on a large dataset and provides well-established feature extraction capabilities.

EfficientNet-B4 is one of the intermediate models in the EfficientNet family, falling between the smaller models (B0-B3) and the larger models (B5-B7). It strikes a balance between model size and accuracy, making it suitable for a wide range of applications [59].

5.3.1 Architecture

The architecture of EfficientNet-B4 builds upon several core concepts:

- **Compound Scaling**: EfficientNet-B4 employs a compound scaling strategy that scales up the network's depth, width, and input resolution in a balanced way. This approach allows the model to achieve high accuracy without excessively increasing computational costs.

- **Mobile Inverted Bottleneck Convolution (MBConv)**: EfficientNet-B4 uses MBConv blocks, which are a type of depthwise separable convolution designed to improve computational efficiency. These blocks help the model capture complex features while minimizing the number of parameters and operations.

- **Swish Activation Function**: EfficientNet-B4 utilizes the Swish activation function, which has been shown to outperform traditional activation functions like

ReLU in various tasks. Swish is defined as $f(x) = x \cdot \text{sigmoid}(x)$, and it helps in improving the model's accuracy and convergence.

5.3.2 Performance

EfficientNet-B4 offers a favorable trade-off between accuracy and computational efficiency. It provides state-of-the-art performance on image classification tasks while requiring fewer resources compared to larger models. The model is particularly well-suited for scenarios where computational resources are limited but high accuracy is still required.

5.3.3 Application in PCA

In the context of Principal Component Analysis (PCA), EfficientNet-B4 is used to extract features from images. Specifically, we leverage the output from the last fully connected layer of pretrained EfficientNet-B4 as input to our PCA process. This feature representation captures the high-level characteristics of the facial images, which are then transformed into principal components for further analysis.

EfficientNet-B4's efficient architecture ensures that the extracted features are both rich and computationally manageable, providing a solid foundation for dimensionality reduction and subsequent machine learning tasks.

5.4 Principal Component Analysis (PCA)

PCA is applied to each segmented face part to enhance feature separation while preserving key characteristics of the data. The input to PCA is taken from the last fully connected layer of the EfficientNet-B4 model, with the process tailored for each facial region: eyes, nose, and mouth.

For each facial part, PCA is used to extract principal components from the feature maps. The number of components is selected based on the performance of the PCA in separating the features:

- For the eyes and nose regions, we extract 1997 principal components, as this num-

ber provided optimal separation of features.

- For the mouth region, we use 143 principal components, which captures 95

5.5 Multi-Layer Perceptrons (MLPs)

Multiple MLPs are trained on the PCA-transformed data of each face part. The MLP architecture includes:

- **Input Layer**: Matches the number of principal components.
- **Output Layer**: A single neuron with a sigmoid activation function for binary classification.

For the training of the Multi-Layer Perceptron (MLP) classifier, we start by using the original face images, along with their corresponding masks and landmarks. After applying the necessary manipulations and blending to these images, we then extract the relevant PCA components for each facial part.

5.6 Performance Metrics

To evaluate the performance of our facial image classification model, we utilize two primary metrics: **Accuracy** and **Area Under the Curve (AUC)**. These metrics help us understand how well the model performs in classifying facial features and distinguishing between different classes.

5.6.1 Accuracy

Accuracy is a fundamental metric that measures the proportion of correctly classified instances among the total number of instances. It provides a general indication of the model's performance across all classes. The formula for calculating accuracy is:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$$

Accuracy is useful for assessing how well the model performs overall. However, it

may not fully reflect the model's performance, especially in cases where the dataset is imbalanced. In such scenarios, the model may perform well on the majority class but poorly on the minority class, which accuracy alone might not reveal.

5.6.2 Area Under the Curve (AUC)

The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) is a performance metric that evaluates the model's ability to distinguish between positive and negative classes. The ROC curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) across various threshold settings. The AUC represents the area under this curve and provides an aggregate measure of performance across all possible thresholds.

The AUC value ranges from 0 to 1:

- **AUC = 1**: Indicates perfect model performance, where the model can perfectly distinguish between all positive and negative instances.
- **AUC = 0.5**: Indicates no discriminative power, which is equivalent to random guessing.
- **AUC < 0.5**: Suggests that the model performs worse than random guessing, indicating potential issues with the model or data.

AUC is particularly useful in evaluating models with imbalanced datasets, where some classes may be underrepresented. It provides a comprehensive view of the model's performance across various threshold levels, offering insights into its ability to classify instances correctly in different scenarios.

5.7 Summarry

The integration of Explainable AI allows for transparency in classification decisions. By analyzing the impact of different facial masks and identifying altered facial regions, the model provides insights into how specific parts of the face influence the final classification outcome. This explanation helps in understanding the modifications made to create fake faces and enhances the interpretability of the facial analysis process.

The process is visually summarized in Figure **??**, which illustrates the complete training phase from data preparation through PCA transformation to MLP training.

6

Experiments and Results

6.1 Introduction

This chapter provides the outcomes of our research, which concentrate on using sophisticated deep learning methods. The main goal was to assess the efficacy of the suggested techniques in producing and detecting deepfake through deliberate alterations and thorough machine learning evaluations. In addition, our objective was to create interpretable AI models that not only achieve high accuracy in detecting deepfakes but also offer transparent and comprehensible explanations of the decision-making process.

Our methodology entails the utilization of masks to target precise facial regions, namely the nose, mouth, and eyes, enabling deliberate alterations in these specific locations. Subsequently, the altered areas were reintegrated into the initial photos, so generating edited. By employing this technique, we were able to produce focused deepfake alterations that highlight specific facial characteristics, which are frequently significant signs of tampering.

In order to examine these photographs, we employed EfficientNet, an advanced convolutional neural network, to assess both the unaltered and modified images. Principal Component Analysis (PCA) representations were generated by extracting features from the layer that comes before the final fully linked layer. The PCA features capture the most important differences in the data, which are essential for differentiating between genuine and counterfeit photos.

Furthermore, a notable component of our work involved the creation of interpretable artificial intelligence models. By utilizing advanced methods such as GradCAM++, our goal was to not only achieve a high level of accuracy in detecting deepfakes but also to offer a clear comprehension of the decision-making process of the model. The models we developed for distinct regions of the face, known as "Eyes," "Nose," and "Mouth,"

were designed to concentrate on specific facial locations. By utilizing explainable AI methodologies, we were able to visually represent the regions of interest that the models relied on for classification. This methodology guarantees that our algorithms are not only efficient in detecting but also explainable, providing distinct insights into which facial characteristics are most indicative of manipulation.

Our study extends beyond simple categorization and makes a valuable contribution to the field of explainable AI. We utilize our models to clarify the fundamental mechanisms involved in deepfake detection. By incorporating explainable AI methodologies, we improve the clarity and reliability of our models, rendering them significant assets not only for identifying deepfakes but also for comprehending the decision-making process of AI systems.

For the final classification task, a Multi-Layer Perceptron (MLP) was trained to differentiate between original and manipulated images based on the PCA features. The results are presented in two main sections:

- **Comparison with Existing Methods**: Evaluating the performance of our approach against state-of-the-art techniques, offering a comprehensive analysis of its strengths and potential areas for improvement.
- **Explainable artificial intelligence**: Exploring the interpretability of our model's decisions, with insights into which facial regions and features are most influential in determining whether an image is real or fake.

6.2 Comparison with Existing Methods

In this section, we compare the performance of our proposed method with that of various existing models across different datasets. We evaluate how well each approach distinguishes between real and manipulated facial images, focusing on both individual facial components (eyes, nose, mouth) and ensemble methods that combine these features to enhance classification performance.

The performance analysis of our models across various datasets reveals their strong capabilities, particularly in deepfake detection on the Celeb-DF (CDF), DFDC, and FFIW datasets. The results of the experiments are presented in Table 6.1. Our "Mouth"

model outperforms the previous state-of-the-art EFNB4 + SBIs model on the Celeb-DF dataset, achieving the highest AUC of **93.85%**. This surpasses the EFNB4 + SBIs model, which had an AUC of 93.18%. Additionally, our "Nose" and "Eyes" models also perform competitively on Celeb-DF, with AUCs of **93.22%** and 92.63%, respectively. This indicates that focusing on specific facial regions, particularly the mouth, can significantly enhance deepfake detection accuracy in challenging scenarios like Celeb-DF.

In the DFDC dataset, our models also demonstrate strong performance. The "Eyes" model achieves an AUC of 71.29%, the "Nose" model 71.12%, and the "Mouth" model 70.74%. Although these results are slightly lower than the EFNB4 + SBIs model, which achieved an AUC of **72.42%**, they remain competitive. This suggests that while region-specific models may excel in certain scenarios, there can be trade-offs in others, indicating that further fine-tuning or hybrid approaches could improve performance in such complex datasets.

Table 6.1: Comparison of AUC Scores Across Different Datasets

Method	**CDF**	**DFDC**	**FFIW**
DSP-FWA [65]	69.30	-	-
LRL [66]	78.26	-	-
FRDM [67]	79.4	-	-
PCL + I2G [68]	90.03	67.52	-
Two-branch [69]	76.65	-	-
DAM [70]	75.3	-	-
LipForensics [71]	82.4	-	-
FTCN [72]	86.9	71.00	74.47
EFNB4 + SBIs[7]	93.18	**72.42**	**84.83**
Eyes (Ours)	92.63	71.30	80.56
Nose (Ours)	**93.22**	71.12	**83.21**
Mouth (Ours)	**93.85**	70.74	**83.30**

On the FFIW dataset, our models again show strong results, particularly with the "Mouth" and "Nose" models, which achieve AUCs of **83.30%** and **83.21%**, respectively. These results, while slightly below the EFNB4 + SBIs model (**84.83%**), still outperform several other state-of-the-art methods, including FTCN, which achieved 74.47%, and Face X-ray + BI, which had 71.15%. The "Eyes" model also performs well on FFIW, with an AUC of 80.56%, further highlighting the effectiveness of region-specific approaches in deepfake detection.

Overall, these results suggest that our models are highly effective, particularly in the Celeb-DF and FFIW datasets. The "Mouth" model, in particular, demonstrates superior performance, indicating that concentrating on specific facial regions can be a powerful strategy in deepfake detection. However, the slight underperformance on the DFDC dataset points to the potential benefits of further refining these models or exploring hybrid methods to enhance their robustness across different datasets.

In the next experiment, we analyze the performance of our local models—focusing on the eyes, nose, and mouth—across various datasets. We explore whether combining the decision outputs of these different local experts can improve the results. To this end, we evaluate several score and decision-level fusion techniques designed to integrate the outputs of the Local Feature Analysis with Deepfake Detection Model (LFA-SBI) models for the eyes, nose, and mouth.

Fusion techniques employed include the Maximum Ensemble, Minimum Ensemble, Voting Ensemble, Median Ensemble, Geometric Mean Ensemble, Harmonic Mean Ensemble, and Weighted Average Ensemble. Each technique combines predictions from the individual models—face, nose, mouth, and eyes—in different ways to determine the final output.

The Maximum Ensemble method, which integrates predictions from all models, achieved the highest performance on the FF++ dataset, with an accuracy of **0.9757**. This method averages the results from the face, nose, mouth, and eyes models, providing a robust combination that maximizes performance across the dataset. In contrast, other fusion techniques like the Minimum Ensemble or Voting Ensemble showed varied performance, with accuracies of **0.672** and **0.700** on the FFIW dataset, respectively.

These results suggest that while combining outputs from multiple models generally enhances detection accuracy, the choice of fusion technique can significantly impact performance. Specifically, the Maximum Ensemble consistently outperforms other techniques, highlighting its effectiveness in leveraging the strengths of individual models. This analysis underscores the potential of ensemble methods in improving deepfake detection and the importance of selecting appropriate fusion strategies for different datasets.

The analysis of accuracy across multiple datasets (FFIW, CDF, FF++, DFDC) reveals that the effectiveness of models can vary significantly depending on the dataset and

Table 6.2: Accuracy Comparison of Different Methods Across Datasets

Method	FFIW	CDF	FF++	DFDC
Eyes	0.692	0.8456	**0.9714**	**0.6093**
Nose	0.702	0.8494	0.9671	0.6029
Mouth	**0.712**	**0.8591**	0.9657	0.5985
SBI	0.684	0.8552	0.9700	0.6081
Maximum Ensemble	0.712	0.8533	**0.9757**	**0.6175**
Minimum Ensemble	0.672	0.8514	0.9600	0.5911
Voting Ensemble	0.700	0.8514	0.9700	0.6033
Median Ensemble	0.706	0.8533	0.9700	0.6047
Geometric Mean Ensemble	0.702	0.8533	0.9700	0.6029
Harmonic Mean Ensemble	0.700	0.8552	0.9686	0.6013
Weighted Average Ensemble	0.700	0.8514	0.9700	0.6033

the specific facial region being analyzed. Notably, the "Mouth" model consistently outperforms other individual models on the FFIW and CDF datasets, achieving accuracy scores of **0.712** and **0.8591**, respectively. These results suggest that the mouth region plays a crucial role in distinguishing between real and manipulated images, indicating its importance in deepfake detection tasks. The consistently superior performance of the "Mouth" model across these diverse datasets underscores the significance of this facial region in enhancing model robustness and accuracy.

Among the models in this experiment, the Maximum Ensemble, which integrates the outputs from the face, nose, mouth, and eyes models, achieved the highest accuracy of **0.9757** on the FF++ dataset. However, the "Eyes" model has the highest individual accuracy at **0.9714**. Upon comparison with other datasets, it is apparent that the DFDC dataset is more challenging, which leads to a decrease in the accuracy of all models. The "Eyes" model outperforms the individual models with an accuracy of **0.6093**, while the Maximum Ensemble model maintains its position as the most accurate at **0.6175**. The SBI model demonstrates consistent performance across all datasets; however, it is somewhat outperformed by the Maximum Ensemble or the top-performing individual models. This suggests that although the SBI model is a reliable solo strategy, integrating multiple models typically leads to superior outcomes. The Maximum Ensemble technique typically offers the most resilient solution across various datasets, indicating that the integration of multiple models may be more successful in reaching greater accuracy, particularly on more demanding datasets such as DFDC (Table 6.2).

6.3 Explainable artificial intelligence

6.3.1 Visualization of Model Focus Using Heatmaps

To further investigate the effectiveness of our region-specific models, we visualize the attention focus of each model using heatmaps. Figure 6.1 displays the heatmaps generated by the "Eyes," "Nose," and "Mouth" models compared to a baseline model trained on the whole face. Each row in the figure corresponds to a specific model, and the columns show the input face image with the region of interest and the corresponding heatmap. Figure 6.1 illustrates the differences in attention focus between our region-specific models and a baseline model trained on the whole face. The baseline model's heatmap, shown in the last row, reveals that its attention is not effectively concentrated on the face. Instead, the attention is dispersed on one part of the image, indicating a lack of focus on the regions that are most critical for detecting deepfake artifacts.

In contrast, the heatmaps generated by our "Eyes," "Nose," and "Mouth" models demonstrate significantly improved attention focus. The "Eyes" model, for instance, directs its attention almost exclusively to the eye regions, which are key for identifying inconsistencies in blinking patterns and gaze direction—common indicators of deepfake manipulations. The "Nose" model, as expected, concentrates on the nose area, which, while less frequently manipulated, can still reveal subtle discrepancies. Finally, the "Mouth" model focuses strongly on the mouth region, which is often challenging to accurately synthesize, particularly in terms of lip synchronization with speech.

The analysis presented in Figure 6.1 highlights the advantages of our region-specific models over the whole-face baseline. By forcing the model to focus on specific facial regions, we ensure that the attention is directed towards areas where deepfake artifacts are most likely to occur. This focused attention is particularly beneficial in detecting high-quality deepfakes, where artifacts may be subtle and localized.

Moreover, the effectiveness of the "Eyes," "Nose," and "Mouth" models in isolating and analyzing specific regions suggests that a composite approach, where these region-specific models are integrated into a broader detection framework, could enhance overall detection performance. Future research could explore such hybrid models that combine

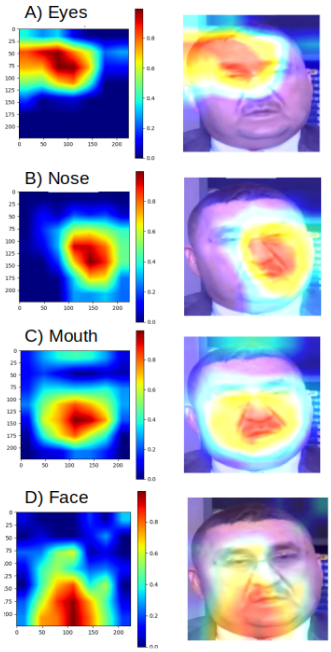

Figure 6.1: Heatmaps of the attention focus of different models. The first column shows the input face image with the region of interest highlighted, while the second column shows the corresponding heatmap. The baseline model's heatmap (top row) does not focus on the face effectively, while our region-specific models ("Eyes," "Nose," and "Mouth") correctly focus on the respective facial regions.

the strengths of each region-specific model, potentially leading to even higher accuracy in deepfake detection across diverse datasets.

6.3.2 Confidence Analysis and Model Behavior

In addition to the primary findings of our research, we further analyzed the output by extracting the most confident predictions from our models for different facial regions across various datasets. which means the model which produced the highest score among all local models. This analysis aimed to understand which facial regions contributed the most to the detection of deepfakes and how this varied across different datasets.

Table 6.3 summarizes the number of most confident detections for each facial region across three datasets: FFIW, FF++, and CDF.

Table 6.3: Confidence Analysis Across Different Datasets and Facial Regions

Facial Region	FFIW	FF++	CDF
Eyes	335	245	168
Nose	60	167	2
Mouth	97	173	265
Face	8	115	86

Among all datasets, the mouth region produced the greatest number of highly confident detections—97 for FFIW, 173 for FF++, and 265 for CDF. This totals to 535

confident detections, indicating that the mouth region has a significant impact on deepfake detection. The mouth region likely shows more prominent manipulations or is more easily detectable due to the dynamic nature of facial expressions, which could be a key focus for most deepfake algorithms.

The nose region showed varying numbers of confident detections. It had 60 detections in FFIW, 167 in FF++, and just 2 in CDF. The relatively higher number in FF++ suggests that nose manipulations are more pronounced in this dataset, or the dataset contains more subtle features that the model can detect in the nose area. However, the extremely low count in CDF indicates that either nose manipulations are less common or less detectable in this dataset.

The eyes region had a moderate number of confident detections compared to the mouth, with 335 for FFIW, 245 for FF++, and 168 for CDF. This suggests that while the eyes are important, they might not be as distinctively manipulated or as easily detectable as the mouth region.

The face region, which encompasses the entire face, had 8 detections in FFIW, 115 in FF++, and 86 in CDF. This suggests that holistic facial manipulations are either subtler or less detectable by the models compared to more localized regions like the mouth or eyes.

The FF++ dataset, eyes had the highest total number of confident detections, particularly in the nose and face regions. This suggests that this dataset might include more varied or pronounced facial manipulations, making it easier for the models to detect specific regions.

In contrast, the CDF dataset had the highest number of confident detections in the mouth region, indicating that the model detects manipulations in the mouth area more effectively in this dataset. This could be due to the nature of the manipulations or the dataset's specific characteristics.

6.3.3 Visualization of Model Confidence and Its Implications for Explainable AI

Furthermore, to better understand and visualize the model's behavior, we generated heatmaps based on the confidence scores for each image across different datasets. These heatmaps reveal the regions where the model's confidence is highest, with more intense

colors indicating higher confidence.

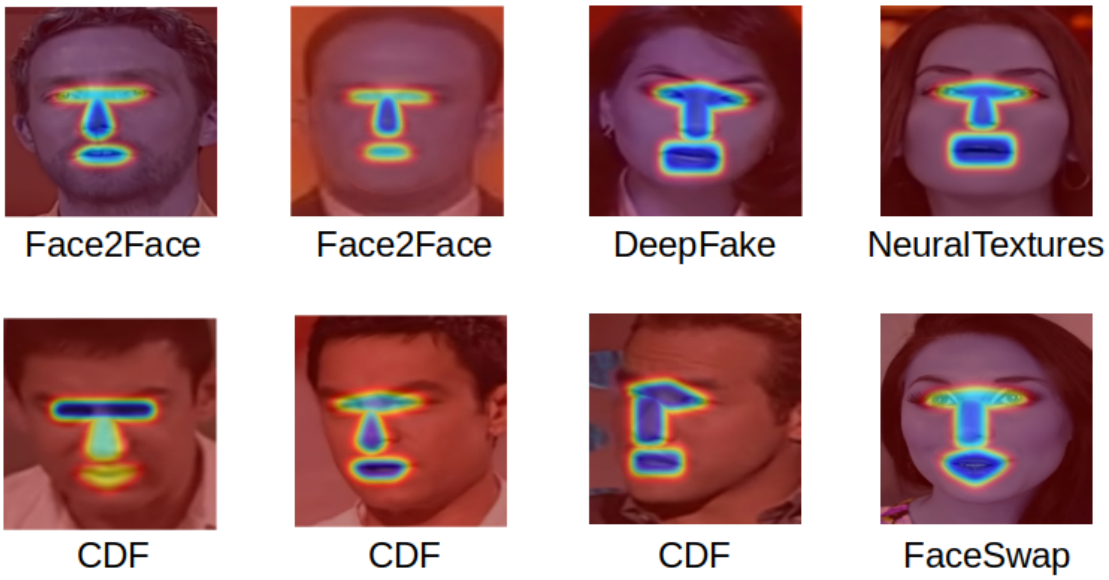

Figure 6.2: Heatmap visualizations of model confidence across different datasets. Blue areas indicate regions of higher confidence.

The heatmap visualizations offer further insights into the variability and strengths of our model's predictions.

Face2Face

In the Face2Face dataset, the model shows a strong focus on the eye and nose regions, as indicated by the high confidence (blue areas). This suggests that the manipulations in these regions are more detectable in this dataset, possibly due to artifacts introduced during the face-swapping process6.2.

DeepFake

The heatmaps for the DeepFake dataset show high confidence around the nose and mouth regions. The model's emphasis on these areas could be due to the nature of manipulations in the dataset, which may involve more pronounced alterations in the central facial features.

NeuralTextures

For NeuralTextures, the model appears to focus significantly on the nose and mouth, with some attention to the eye region. This distribution of confidence might reflect the

specific manipulation techniques used in this dataset, which may cause more detectable distortions in these areas.

FaceSwap

The heatmaps for the FaceSwap dataset show the highest confidence in the mouth region, with significant attention also given to the eyes and nose. This pattern suggests that the FaceSwap manipulations may particularly affect the dynamic areas of the face, such as the mouth, making these regions crucial for detection.

CDF (Celeb-DF)

Interestingly, the heatmaps for the CDF dataset show a more varied distribution of attention across different facial regions. Unlike other datasets, where the model's focus is more consistent, the CDF heatmaps reveal that the model adapts its attention based on the specific manipulations present in each image. For example, some images show high confidence in the nose region, while others highlight the mouth or eyes. This variability indicates that the model is not only sensitive to a wide range of manipulations but is also capable of adjusting its detection focus based on the unique characteristics of each image.

Overall, the analysis of confident detections across facial regions and datasets underscores the significance of the mouth region in deepfake detection, with notable contributions from the eyes and nose. The variability observed across datasets suggests that the effectiveness of deepfake detection might be dataset-dependent and that different datasets may highlight different facial manipulation characteristics.

6.3.4 Analysis of Generalization Across Datasets

The table presents the AUC scores of models trained on the FF++ dataset and evaluated on both FF++ and Celeb-DF (CDF). The models are trained on specific facial regions (eyes, nose, mouth) using different quantities of images ($N \times 16$ and $N \times 32$).

Performance on FF++ (Training Dataset)

- $N \times 16$ **Images**: The models perform exceptionally well on FF++, with AUCs ranging from 0.9980 to 0.9992, indicating that the PCA-MLP pipeline is highly effective when evaluated on the training dataset.
- $N \times 32$ **Images**: A slight decrease in performance is observed with more images, shows the introduction of less relevant features. Nevertheless, the AUC remains very high, confirming the models' strong performance on FF++.

Generalization to CDF (Evaluation Dataset)

- $N \times 16$ **Images**: There is a noticeable drop in AUC when the models are evaluated on CDF, with scores ranging from 0.6942 to 0.7908. This suggests that the features learned from FF++ do not transfer perfectly to CDF, indicating a potential difference in the types of manipulations or video qualities between the two datasets.
- $N \times 32$ **Images**: Increasing the number of training images results in a significant improvement in CDF performance, with AUCs rising to 0.9263 for Eyes, 0.9322 for Nose, and 0.9385 for Mouth. This indicates that a larger and more diverse training set helps the model generalize better to new, unseen data and the fact that the models are capable of highly generalization to other datasets.

Table 6.4: Comparison of AUC Scores Across Quantity of Images

Method	**Number of Images**	**FF++**	**CDF**
Eyes	5760×16	0.9987	0.6942
Nose	5760×16	0.9992	0.7908
Mouth	5760×16	0.9980	0.7279
Eyes	5760×32	0.9961	0.9263
Nose	5760×32	0.9953	0.9322
Mouth	5760×32	0.9944	0.9385

In summary, the results of this experiment can be analysed in three steps:

- **Model Generalization Capability**: The AUC scores reveal that while all models achieve near-perfect performance on the FF++ dataset, there is a notable decline

when these models are applied to the Celeb-DF (CDF) dataset. This suggests that while the models are highly effective in detecting deepfakes within the domain they were trained on, their ability to generalize to different datasets with varying characteristics is more limited. This generalization gap highlights the challenge of ensuring that models can maintain their effectiveness across diverse datasets.

- **Model Adaptability with Increased Training Data**: The results show that increasing the number of training images (5760×32) significantly enhances the generalization performance of the models, particularly on the CDF dataset. This improvement underscores the adaptability of the models, suggesting that when provided with more extensive and diverse training data, these models are capable of learning more robust and transferable features. However, the Nose model stands out in its ability to generalize even with fewer images, indicating a higher intrinsic robustness of the features it extracts from the nose region.

- **Nose Model Robustness**: Among the models, the Nose model consistently demonstrates strong performance across both FF++ and CDF datasets, outperforming the Eyes and Mouth models, especially in terms of generalization. This suggests that the features extracted from the nose region by this model are inherently more transferable across different deepfake detection scenarios, making it a particularly valuable approach for developing more generalized deepfake detection systems. The success of the Nose model implies that certain facial regions may inherently capture more relevant features for deepfake detection, particularly when models are evaluated on datasets that differ from the training data.

These findings underscore the importance of considering dataset diversity and the quantity of training data when developing deepfake detection models, as well as the potential challenges of generalizing from one dataset to another.

6.3.5 Deepfake Detection Performance Based on Facial Regions

In this experiment, we tested the models across four different types of deepfake manipulations: DeepFake, Face2Face, FaceSwap, and NeuralTextures. The performance of each model was measured using the Area Under the Curve (AUC) and Accuracy metrics, as

summarized in Table 6.5.

Table 6.5: Performance of Deepfake Detection Models on Different Facial Regions

Region	DeepFake		Face2Face		FaceSwap		NeuralTextures	
	AUC	**Accuracy**	**AUC**	**Accuracy**	**AUC**	**Accuracy**	**AUC**	**Accuracy**
Nose	0.9998	0.9892	0.9974	0.9857	0.9984	0.9785	0.9846	0.9428
Eyes	0.9997	0.9892	0.9983	0.9821	0.9984	0.9750	0.9869	0.9500
Mouth	0.9998	0.9892	0.9960	0.9821	0.9987	0.9857	0.9813	0.9321
Face	0.99098	0.9928	0.9984	0.9821	0.9989	0.9785	0.9856	0.9464

The results in Table 6.5 demonstrate the effectiveness of deepfake detection models when focusing on different facial regions, with performance varying across the different manipulation types. The model focusing on the nose region achieved excellent performance, particularly with the DeepFake and Face2Face manipulations, where the AUC values were nearly perfect. This suggests that the nose provides strong discriminative features, likely due to its relatively stable and distinct shape in comparison to other facial regions. The eyes region also yielded high AUC values across all manipulation types, indicating that eye features are critical for detecting deepfakes, although the accuracy for FaceSwap and NeuralTextures was slightly lower, possibly because these techniques can more effectively manipulate or blend eye movements to mimic natural behavior. The mouth region showed the highest accuracy for FaceSwap, which is expected since FaceSwap often targets the mouth region to match the source face, making it more susceptible to manipulation, hence easier to detect. Finally, the model analyzing the entire face consistently performed well across all metrics and manipulation types, suggesting that while individual regions are important, a holistic approach that considers the entire face captures a broader range of features, leading to improved detection performance. The results underscore the importance of understanding which facial regions are manipulated by different deepfake techniques, as this knowledge can significantly influence the effectiveness of detection models.

7

Conclusions

This work has demonstrated the impact of individual facial components, such as the eyes, nose, and mouth, on the overall accuracy of deepfake identification by dividing the face into separate areas and applying certain masks.

The utilization of Principal Component Analysis (PCA) has been essential in this research. PCA effectively reduced noise and identified the main components of each facial region, hence enabling more precise and meaningful feature extraction. PCA has enhanced the capacity to differentiate between distinct facial characteristics by decreasing their dimensionality and emphasizing the most significant alterations. This technique has improved the ability of MLP models to differentiate between authentic and manipulated photographs by giving priority to the most crucial features in each facial region.

The results emphasize the importance of the oral region in detecting deepfakes, as it consistently generated the highest number of confident identifications across different datasets. The nose and eyes regions also played a role in the classification, but with a relatively insignificant influence. The individualized masks employed for each specific site demonstrated the capacity to concentrate on crucial facial characteristics in order to augment the precision of detection, thus affirming the effectiveness of our approach.

Moreover, the analysis of heatmap visualizations has revealed the model's adaptability across different datasets, particularly in the Celeb-DF (CDF) dataset. The heatmaps indicated that the model adjusts its focus depending on the specific manipulations present in each image, effectively detecting subtle alterations in different facial regions. This adaptability underscores the power of using confidence-based heatmaps as a tool for Explainable AI (XAI), enhancing the transparency and interpretability of deepfake detection models. The ability to visualize and analyze confidence distributions across facial regions not only improves model explainability but also highlights the potential to refine

and tailor detection methods to specific types of manipulations, as observed in datasets like CDF.

This study provides valuable insights into the impact of different facial features on classification judgments within the Explainable AI framework. The work improves our understanding of the model's decision-making process by elucidating the facial regions that exert the greatest influence. AI systems must be open in order to fully comprehend model behavior and develop confidence, especially in sensitive applications such as deepfake detection.

Furthermore, the incorporation of data augmentation and blending techniques has enhanced the robustness of the models, allowing them to proficiently handle diverse distortions and alterations in facial images. Using EfficientNet-B4 for feature extraction allowed for a balance between efficiency and efficacy, making it possible to capture important facial characteristics while efficiently utilizing computer resources.

This study advances the field of facial image classification and deepfake detection by demonstrating the efficacy of segmenting face features and employing Principal Component Analysis (PCA) to improve the model's performance. Moreover, it enriches the domain of Explainable Artificial Intelligence (XAI) by offering significant insights into the specific facial characteristics that influence the process of forming categorization judgments. The findings, particularly those related to the adaptable focus of the model in the CDF dataset, provide a foundation for future research in the areas of facial recognition and deepfake detection, emphasizing the significance of further exploring the impact of different facial components on feature analysis and model interpretability.

Bibliography

[1] B. U. Mahmud and A. Sharmin, "Deep insights of deepfake technology: A review," *arXiv preprint arXiv:2105.00192*, 2021.

[2] D. Güera and E. J. Delp, "Deepfake video detection using recurrent neural networks," in *2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS)*. IEEE, 2018, pp. 1–6.

[3] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-df: A large-scale challenging dataset for deepfake forensics," pp. 3207–3216, 2020.

[4] T. Zhou, W. Wang, Z. Liang, and J. Shen, "Supplemental material: Face forensics in the wild," 2021, eTH Zurich, Beijing Institute of Technology, Inception Institute of Artificial Intelligence. [Online]. Available: https://github.com/tfzhou/FFIW

[5] N. Bonettini, E. D. Cannas, S. Mandelli, L. Bondi, P. Bestagini, and S. Tubaro, "Video face manipulation detection through ensemble of cnns," in *2020 25th international conference on pattern recognition (ICPR)*. IEEE, 2021, pp. 5012–5019.

[6] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics++: Learning to detect manipulated facial images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1–11.

[7] K. Shiohara and T. Yamasaki, "Detecting deepfakes with self-blended images," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 1816–1826.

[8] A. Rosebrock, "Facial landmarks with dlib, opencv, and python," 2017, accessed: 2024-07-21. [Online]. Available: https://pyimagesearch.com/2017/04/03/facial-landmarks-dlib-opencv-python/

[9] S. Chesney and D. K. Citron, "Deepfakes: A looming challenge for privacy, democracy, and national security," *California Law Review*, vol. 107, no. 6, pp. 1753–1819, 2019.

[10] B. Dolhansky, S. Yang, M. M. Rahaman, L. Chen, and J. Leskovec, "The deepfakes detection challenge and dataset," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. [Online]. Available: https://arxiv.org/abs/2006.07397

[11] A. Rossler, D. Cozzolino, L. Verdoliva, and E. Ricci, "Faceforensics++: Learning to detect manipulated facial images," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. [Online]. Available: https://arxiv.org/abs/1901.08971

[12] Z. Wang, D. Li, J. Dong, and X. Jiang, "A survey of deepfake detection techniques," *IEEE Access*, vol. 9, pp. 17 558–17 578, 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9350329

[13] J. Stehouwer, H. Dang, F. Liu, X. Liu, and A. K. Jain, "On the detection of digital face manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 5781–5790. [Online]. Available: https://openaccess.thecvf.com/content_CVPR_2020/html/Stehouwer_On_the_Detection_of_Digital_Face_Manipulation_CVPR_2020_paper.html

[14] W. H. Abir, F. R. Khanam, K. N. Alam, M. Hadjouni, H. Elmannai, S. Bourouis, R. Dey, and M. M. Khan, "Detecting deepfake images using deep learning techniques and explainable ai methods," *Intelligent Automation Soft Computing*, vol. 35, no. 2, pp. 2151–2169, 2023. [Online]. Available: https://doi.org/10.32604/iasc.2023.029653

[15] Z. Sun, S. Hu, Z. Lin, H. Sun, Z. Wu, and C. Wang, "Fst-matching: Exploiting fine-grained spatial-temporal information for deepfake detection," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 3124–3132. [Online]. Available: https://dl.acm.org/doi/10.1145/3394171.3413678

[16] M. S. Rana and A. H. Sung, "Deepfakestack: A deep ensemble-based learning technique for deepfake detection," in *2020 IEEE International Conference on Cloud Computing and Edge Computing (CSCloud-EdgeCom)*. IEEE, 2020, pp. 97–104. [Online]. Available: https://doi.org/10.1109/cscloud-edgecom49738.2020.00021

[17] R. Plesh, J. Krizaj, K. Bahmani, M. Banavar, V. Štruc, and S. Schuckers, "Discovering interpretable feature directions in the embedding space of face recognition models."

[18] Authors, "Discovering interpretable feature directions in the embedding space of face recognition models," *Journal/Conference*, Year.

[19] S. Vyas, "Historical perspectives on media manipulation and deepfake technology," *Journal of Media History*, vol. 8, no. 3, pp. 15–30, 2022.

[20] A. Ovadya, I. Schwartz, C. Popa, D. Gunter, and C. Lee, "The ethics of machine learning," in *Proceedings of the 2018 IEEE Symposium on Security and Privacy*. IEEE, 2018, pp. 123–135.

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

[22] T. Karras, S. Laine, and T. Aila, "Progressive growing of gans for improved quality, stability, and variation," in *International Conference on Learning Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=Hk3L3iAqK7

[23] ——, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4401–4410.

[24] ——, "Analyzing and improving the image quality of stylegan," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8110–8119, 2020. [Online]. Available: https://openaccess.thecvf.com/content_CVPR_2020/html/Karras_Analyzing_and_Improving_the_Image_Quality_of_StyleGAN_CVPR_2020_paper.html

[25] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3d faces," in *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, 1999, pp. 187–194.

[26] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2face: Real-time face capture and reenactment of rgb videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2387–2395.

[27] C. Richardt, A. Richter, J. M., and A. T., "Neural textures: Combining neural networks with graphics rendering," *Journal of Computer Graphics Techniques*, 2017. [Online]. Available: https://www.jcgt.org/published/0001/02/01

[28] O. Fried, Y. Sheikh, and A. M., "Texturizing videos with neural networks," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 123–132, 2018.

[29] Y. Shen, P. Luo, J. Yan, X. Wang, and X. Tang, "Faceid-gan: Learning a symmetry three-player gan for identity-preserving face synthesis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2018, pp. 821–830.

[30] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8789–8797.

[31] T. Xiao, J. Hong, and J. Ma, "Elegant: Exchanging latent encodings with gan for transferring multiple face attributes," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 168–184.

[32] M. Liu, M. Zhang, Z. Liu, Z. Wang, T. Zhang, J. Han, and P. Liu, "Stgan: A unified selective transfer network for arbitrary image attribute editing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3651–3660.

[33] T. Li, R. Qian, C. Dong, S. Liu, Q. Yan, W. Zhu, and L. Lin, "Beautygan: Instance-level facial makeup transfer with deep generative adversarial network," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 645–653.

[34] I. Korshunova, W. Shi, J. Dambre, and L. Theis, "Fast face-swap using convolutional neural networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017, pp. 3677–3685.

[35] R. Natsume, T. Yatagawa, and S. Morishima, "Fsnet: An identity-aware generative model for image-based face swapping," in *Asian Conference on Computer Vision (ACCV)*. Springer, 2018, pp. 117–132.

[36] Q. Sun, A. Tewari, W. Xu, M. Fritz, C. Theobalt, and B. Schiele, "A hybrid model for identity obfuscation by face replacement," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2018, pp. 658–674.

[37] D. Bitouk, I. Matthews, S. Lazebnik, Y. Weiss, and T. Kanade, "Face swapping: A novel approach to face verification and identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Anchorage, AK, USA: IEEE, 2008, pp. 1–8.

[38] L. Verdoliva, "Media forensics and deepfakes: An overview," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 5, pp. 910–932, 2020.

[39] F. Matern, C. Riess, and M. Stamminger, "Exploiting visual artifacts to expose deepfakes and face manipulations," in *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*. IEEE, 2019, pp. 83–92.

[40] A. Sengur, Y. Guo, Y. Akbulut, and W. Zhu, "Hybrid model for deepfake detection: Convolutional neural network and residual-based features," *Applied Soft Computing*, vol. 70, pp. 138–147, 2018.

[41] A. Sengur *et al.*, "Investigation of comparison on modified cnn techniques to classify fake face in deepfake videos," *Journal of Computer and Communications*, vol. 6, no. 6, pp. 21–33, 2018.

[42] S. Tariq, S. Lee, H. Kim, Y. Shin, and S. S. Woo, "Detecting both machine and human created fake face images in the wild," in *Proceedings of the 2nd International Workshop on Multimedia Privacy and Security*, 2018.

[43] N. S. Ivanov, A. V. Arzhskov, and V. G. Ivanenko, "Combining deep learning and super-resolution algorithms for deep fake detection," in *IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)*. IEEE, 2020, pp. 326–328.

[44] A. Das and L. Sebastian, "A comparative analysis and study of a fast parallel cnn based deepfake video detection model with feature selection (fpcdfm)," in *2023 Advanced Computing and Communication Technologies for High Performance Applications (ACCTHPA)*. IEEE, 2023, pp. 1–9.

[45] C.-C. Hsu, Y.-X. Zhuang, and C.-Y. Lee, "Deep fake image detection based on pairwise learning," *Applied Sciences*, vol. 10, no. 2, p. 370, 2020.

[46] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2016, pp. 2818–2826.

[47] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "Emotions don't lie: An audio-visual deepfake detection method using affective cues," in *Proceedings of the 28th ACM International Conference on Multimedia*. ACM, 2020, pp. 2823–2832.

[48] C. M.-C. Li, Y. and S. Lyu, "In ictu oculi: Exposing ai generated fake face videos by detecting eye blinking," in *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE, 2018, pp. 1–7.

[49] D. I. Ciftci, U.A. and L. Yin, "How do the hearts of deep fakes beat? deep fake source detection via interpreting residuals with biological signals," in *2020 IEEE International Joint Conference on Biometrics (IJCB)*, 2020, pp. 1–10.

[50] J. Zhang, B. Dong, Z. Wang, and X. Wang, "Automated face swapping and its detection," *Journal of Visual Communication and Image Representation*, vol. 48, pp. 41–52, 2017.

[51] S. Zhang, Y. Li, H. Qi, and S. Lyu, "Ffiw: Face forensics in the wild," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2021.

[52] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," *International Conference on Machine Learning*, pp. 1597–1607, 2020.

[53] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1422–1430.

[54] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[55] L. Jing, Y. Liu, D. Dong, L. Su, Z. Wu, and Y. Tian, "Self-supervised spatiotemporal representation learning by video pace prediction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 8, pp. 2730–2742, 2020.

[56] D. E. King, "Dlib-ml: A machine learning toolkit," 2009, journal of Machine Learning Research 10 (2009) 1755-1758. [Online]. Available: http://dlib.net

[57] J. Deng, J. Guo, Y. Wen, Z. Li, and W. Liu, "Retinaface: Single-stage dense face localization in the wild," *arXiv preprint arXiv:1905.00641*, 2019. [Online]. Available: https://arxiv.org/abs/1905.00641

[58] "Facial landmarks shape predictor," https://github.com/codeniko/shape_predictor_81_face_landmarks, accessed: 2021-11-13.

[59] P. Zhang, L. Yang, and D. Li, "Efficientnet-b4-ranger: A novel method for greenhouse cucumber disease recognition under natural complex environment," *Computers and Electronics in Agriculture*, vol. 176, p. 105652, 2020. [Online]. Available: https://doi.org/10.1016/j.compag.2020.105652

[60] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009, pp. 248–255.

[61] "Faceswap," https://github.com/MarekKowalski/FaceSwap/, accessed: 2021-11-13.

[62] P. Zhou, X. Xie, Y. Dong, D. Li, L. Wang, X. Chen, and X. Zhang, "Two-stream neural networks for tampered face detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2017/html/Zhou_Two-Stream_Neural_Networks_CVPR_2017_paper.html

[63] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European Conference on Computer Vision (ECCV)*, 2016. [Online]. Available: https://arxiv.org/abs/1512.02325

[64] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 39, no. 6, pp. 1137–1149, 2017. [Online]. Available: https://arxiv.org/abs/1506.01497

[65] Y. Li and S. Lyu, "Exposing deepfake videos by detecting face warping artifacts," in *CVPR Workshops*, 2019, pp. 1–6.

[66] S. Chen, T. Yao, Y. Chen, S. Ding, J. Li, and R. Ji, "Local relation learning for face forgery detection," in *AAAI*, vol. 35, 2021, pp. 1081–1088.

[67] Y. Luo, Y. Zhang, J. Yan, and W. Liu, "Generalizing face forgery detection with high-frequency features," in *CVPR*, 2021, pp. 16 317–16 326.

[68] T. Zhao, X. Xu, M. Xu, H. Ding, Y. Xiong, and W. Xia, "Learning self-consistency for deepfake detection," in *ICCV*, 2021, pp. 15 023–15 033.

[69] I. Masi, A. Killekar, R. M. Mascarenhas, S. P. Gurudatt, and W. AbdAlmageed, "Two-branch recurrent network for isolating deepfakes in videos," in *ECCV*, 2020, pp. 667–684.

[70] T. Zhou, W. Wang, Z. Liang, and J. Shen, "Face forensics in the wild," in *CVPR*, 2021.

[71] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic, "Lips don't lie: A generalisable and robust approach to face forgery detection," in *CVPR*, 2021, pp. 5039–5049.

[72] Y. Zheng, J. Bao, D. Chen, M. Zeng, and F. Wen, "Exploring temporal coherence for more general video face forgery detection," in *ICCV*, 2021, pp. 15 044–15 054.