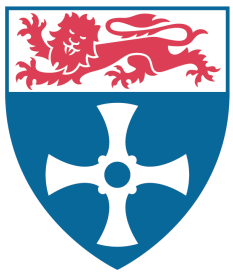
Novel Deepfakes Detection Strategies: Insights from Prosopagnosia



Fatimah Alanazi

School of Computing
Newcastle University

This dissertation is submitted for the degree of
Doctor of Philosophy

October 2024

I would like to dedicate this thesis to my family with love...

Declaration

I declare that this dissertation is an original report of my doctoral research except where specific reference is made to the work of others. This dissertation is written by me and have not been submitted in whole or in part for any previous degree.
Parts of the work presented in this thesis have been published in the following:

1. Alanazi, F. (2022, December). Comparative Analysis of deepfakes Detection Techniques. In 2022 14th International Conference on Computational Intelligence and Communication Networks (CICN) (pp. 119-124). IEEE.

2. Alanazi, F., Morgan, G., Ushaw, G., & Davison, R. Strategies for Addressing Prosopagnosia as a Potential Solution to Facial Deepfake Detection , In the 12th International Conference on Digital Image Processing and Vision (ICDIPV 2023) Vol.13, No.13

3. Alanazi, F., Ushaw, G., & Morgan, G. (2023). Improving Detection of DeepFakes through Facial Region Analysis in Images. Electronics, 13(1), 126.

4. Alanazi, F. (2024). Comparative Analysis of Internal and External Facial Features for Enhanced Deepfake Detection. Presented at the 16th International Conference on Agents and Artificial Intelligence (ICAART 2024).

Fatimah Alanazi
October 2024

Acknowledgements

I extend my deepest gratitude to everyone who contributed to the success of this dissertation.

First and foremost, I owe immense thanks to my supervisors, Prof. Graham Morgan and Dr. Gary Ushaw. Their guidance, encouragement, and wise supervision were crucial to my research success.

Sincere gratitude and thanks are dedicated to my parents for their constant prayers, and to my siblings, brothers, and friends for their support and love.

A heartfelt thank you to my husband Abdullah, and my two precious daughters, Jude and Deem. Your inspiration, motivation, and the boundless joy you have brought to this journey are invaluable. Jude, your enthusiasm and vibrant spirit have been a constant source of inspiration, guiding me through each step of this process. Deem, your ability to fill our home with laughter and light has been a cherished sanctuary and a source of ongoing motivation. The presence of you all in my life is a gift beyond measure, for which I am profoundly grateful.

I extend my deepest gratitude to my husband, Abdullah, for his unwavering love, patience, and understanding. Your support has been a cornerstone in my pursuit of these goals. Your strength and encouragement have been instrumental in navigating the challenges of this journey.

Also,I would like to thank to my country, the Kingdom of Saudi Arabia, for providing all the support I needed to complete my degree.

Lastly, thank you to the wonderful community of Newcastle University for their help during this journey.

Abstract

The credibility of audio and video content, which is essential to our perception of reality, is increasingly challenged by advancements in deepfake generation techniques. Existing detection models primarily focus on identifying anomalies and digital artifacts. However, the rapid evolution of technology enables the creation of sophisticated deepfakes that can evade these methods.

This thesis investigates the effectiveness of different facial features for deepfake detection in images and face recognition in individuals with prosopagnosia. It examines whether there is a correlation between the facial features prioritized by AI models for deepfake detection and those emphasized in training programs aimed at enhancing face recognition in individuals with prosopagnosia. Additionally, it assesses the impact of occluding each facial feature during training on AI model performance and identifies which facial elements individuals with prosopagnosia find most challenging to recognize.

Inspired by research into prosopagnosia, which highlights the importance of internal facial features like the eyes and nose, this study proposes a novel approach to deepfake detection. The methodology involves identifying critical facial features, applying face cut-out techniques to create training images with various occlusions, and evaluating AI models trained on these datasets using EfficientNet-B7 and Xception models.

The results indicate that models trained with occluded datasets performed better, with the EfficientNet-B7 model achieving a higher accuracy rate (92%) when core facial elements (eyes and nose) were covered, compared to models trained on datasets without occlusions or with occlusions covering external features. This suggests that focusing on features outside the face's center improves detection accuracy. The findings also highlight that facial cues beneficial for individuals with prosopagnosia do not uniformly translate to equivalent value for AI models.

This research demonstrates that detection systems can be more effective by focusing on a small region of the face, contributing significantly to the improvement of deepfake detection methods and enhancing our understanding of face recognition processes.

Contents

List of Figures

List of Tables

Nomenclature

Acronyms	Abbreviations
3DMAD	The 3D Mask Attack Database
AAM	Active Appearance Model
AFW	Automatic Face Weighting Algorithm
AI	Artificial Intelligence
ALS	Amyotrophic Lateral Sclerosis
AttGAN	Attribute GAN
CNN	Convolutional Neural Network
CoRNs	Convolutional Reservoir Networks
CP	Congenital Prosopagnosia
CPBD	Cumulative Probability of Blur Detection
DeepFD	Deep Forgery Discriminator
DFT-MF	Discrete Fourier Transform with Matched Filtering
DP	Developmental Prosopagnosia
DRM	Digital Rights Management
DRMF	the Discriminative Response Map Fitting Method
EM	Expectation Maximisation Algorithm
FD2Foremer	Forgery-Detection-with-Facial-Detail Transformer

FR	Facial Recognition
GAN	Generative Adversarial Network
GDWCT	GenerativeDeep Whitening-Coloring Transform
LBP	Local Binary Patterns
LRCN	long-term Recurrent Convolutional Network
LSTM	Long Short-Term Memory
M3ER	Multimodal Emotion Recognition Algorithm
ML	Machine Learning
MTCNN	Multi-task Convolutional Neural Network
PPG	Photoplethysmography
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RNN	Recurrent Neural Network
ROI	Region of Interest
SVM	Support Vector Machine
SWIR	Short-Wave Infraredimaging
TD-3DCNN	Temporal Dropout 3D Convolutional Neural Network
VGG	Visual Graphics Group
VIS	The Visual Light Spectrum

Chapter 1

Introduction

1.1 Overview

In recent years, the realm of Artificial Intelligence (AI) has undergone rapid advancements, leading to its widespread integration across diverse sectors, including engineering, education, and, notably, the medical field. Incorporating AI and intense learning techniques in medical applications has attracted considerable attention for various purposes. These encompass disease diagnosis and the classification of medical imagery. Deep learning has significantly influenced the management of severe medical conditions. A study highlights this [3] that concentrates on employing machine learning methods for the early detection and diagnosis of Alzheimer's disease. Additionally, deep learning is not limited to image analysis but also extends to assisting in the diagnosis of diseases.

The transition in the medical field from conventional methodologies to a growing reliance on AI is evident. Our research focuses on leveraging medical expertise to augment AI capabilities, particularly in advancing deepfake technology.

This research aims to evaluate the correlation between human perception of faces and deepfakes and to understand how deepfakes intersect with medical challenges in facial recognition. With the emergence of automated facial recognition systems, individuals can now be identified across a variety of digital formats, including both still and moving images. This technology has the potential for proactive application in various societal aspects of daily life, from personal identification for security purposes (such as unlocking smartphones) to aiding law enforcement (for instance, in identifying criminals in crowded environments).

Artificial intelligence techniques have become adept at manipulating digital images of individuals and repurposing them for use in moving and still images, a process termed "deepfakes" [4]. Deepfakes allow individuals to appear to be engaging in fabricated activities, which presents significant challenges to society. These challenges arise when people are deceived into erroneously believing that an individual is participating in an activity. This can potentially lead to many undesirable social outcomes, including damage to an individual's reputation, inciting civil unrest, promoting criminal fraud, and even causing people to doubt the authenticity of reality itself [5].

The rise of deepfake technology has created significant challenges for facial recognition services, as deepfakes can deceive their intended audience. The primary issue arises when these systems fail to recognize a face correctly and mistakenly identify deepfakes. This is a crucial area of research because deepfakes present a serious threat. As fake technology advances, it will become increasingly difficult to mitigate the numerous problems associated with false identities [6].

Deepfake analysis has traditionally focused on detecting anomalies within an image (e.g., digital artefacts that may be visible when deepfakes are produced). However, this approach may become unsustainable as technology continues to improve. Therefore, this project proposes a new research direction that utilises medical insights into facial recognition, specifically focusing on a condition known as prosopagnosia.

Prosopagnosia is a condition that prevents individuals from recognizing faces, even those of close family members [7]. This medical condition has been known for many years, and our research explores the existing literature and medical expertise to improve deepfake recognition. We aim to enhance deepfake detection systems by leveraging insights into prosopagnosia, particularly in detecting deepfakes within images.

1.2 Problem Statement

For over a century, audio and video content have played a crucial role in establishing what we consider to be true. In a similar manner, they have been pivotal in shaping our perception of reality. This raises important considerations regarding the reliability of our visual and auditory perceptions in an era where their credibility is often uncertain.

Deepfakes have the potential to manipulate images. Therefore, it is critical to develop and further improve the detection methods used in identifying forged images. However, the majority of existing methods for detecting deepfakes rely on identifying

anomalies commonly found in deepfake images. This is due to the visible digital artefacts that characterise the production process of deepfakes. As technology advances, these digital artefacts become less noticeable, limiting the effectiveness of current detection methods. This limitation arises because more advanced deepfake algorithms produce higher quality and more realistic forgeries that are harder to distinguish from genuine images [8]. Consequently, there is an urgent need for more sophisticated and adaptive detection techniques to keep up with the evolving capabilities of deepfake technology. This dynamic has become an "arms race", with each advance in deepfake creation technology prompting an advance in detection methods, and vice versa.

In the ongoing race between technological innovation and the fight against deepfakes, a noteworthy pattern emerges: as experts identify gaps (lacunae) in the detection mechanisms for fraudulent imagery, malicious actors adeptly exploit these weaknesses to develop enhanced methodologies for generating deepfakes. This scenario underscores a significant challenge in digital content security. The expertise and operational domain of those safeguarding against deepfakes and the perpetrators creating them are markedly similar [5]. Consequently, both parties possess an in-depth understanding of the tactical approaches and nuanced strategies prevalent within this specialised field.

Therefore, the focus of this research is to explore existing literature and expertise in medicine to assist the available technology in deepfake facial recognition, thereby proffering solutions to the challenges posed by deepfakes. The research question revolves around developing a better recognition system gleaned from medical findings on these conditions and/or identifying deepfakes themselves. A comprehensive literature survey on deepfake technology and a review of the literature on the condition of prosopagnosia will be conducted to find the interconnectedness of these two knowledge areas.

Few studies have focused on the association between human perception of faces and deepfakes. Additionally, there are limited studies that examine the link between face recognition disorders and deepfakes. However, there is negligible research that utilises the medical expertise of prosopagnosia to improve deepfake recognition.

1.3 Research Questions

As part of this thesis, we aim to answer the following research questions:
Despite the impressive progress of deepfake detection methods, the creation of deepfakes is still outpacing their ability to detect them. As a result, high-quality deepfakes can

now be produced that are increasingly difficult to distinguish from real images using current technology. This raises serious concerns about the potential for deepfakes to be used to spread misinformation and disinformation.

Question 1: Which facial features are most effective for deepfake detection and face recognition in individuals with prosopagnosia? Is there a correlation between the facial features prioritized by AI models for deepfake detection and those emphasized in training programs aimed at enhancing face recognition in individuals with prosopagnosia?

In Chapter 2, we provide a survey of the literature on deepfake detection and prosopagnosia to identify the overlap between them. The chapter also provides valuable insights into the research conducted in both disciplines. The main goal of this chapter is to explore and identify a connection between deepfake recognition and face recognition disorders.

After identifying the strategies used in the medical field to improve patients' ability to distinguish between faces, we designed our experiment in Chapter 4 based on those strategies. We used the same techniques that have been used in the medical field to test whether a deep neural model could improve its ability to detect the difference between fake and real faces using the same clues from the medical filed.

The findings from both phase one and two lead to support that the operational mechanism of AI models for face recognition differs from the cognitive functioning of individuals with prosopagnosia. Specifically, AI models rely less on the facial cues that are most helpful for people with prosopagnosia.Therefore, we focus on our second question

Question 2: How does the individual occlusion of each facial feature during the training process influence the performance of AI models in detecting deepfakes? What role does each facial feature play in the accuracy and reliability of deepfake detection?

In Chapter 4, specifically, during the third phase of our experiment, we conducted a detailed analysis of individual facial features and their impact on the model's ability to detect deepfakes. Our findings revealed that certain facial regions, such as the nose and mouth, provided less valuable information compared to other features. Notably, the analysis highlighted the eyes as the most critical regions for distinguishing deepfake faces.

Question 3: What specific facial features do individuals with prosopagnosia find most challenging to recognize?How do their perceptions of faces differ from those without the condition?

In Chapter 2, In order to answer our research question, we examined the literature on prosopagnosia. Our review encompassed various dimensions, encompassing the specific attributes prosopagnosia individuals prioritize or avoid when attempting facial recognition. Additionally, we explored the methodologies employed to enhance their face recognition skills. This comprehensive review encompassed sections on face processing, eye movement patterns, fixation behaviors in prosopagnosia patients, prosopagnosia training regimens, and the identification cues employed by prosopagnosia individuals in recognizing faces.

We discovered that patients with prosopagnosia tend to avoid focusing on internal facial features, particularly the eyes, and instead rely more on external facial characteristics and non-facial attributes such as hairstyle and skin tone. However, when these patients were trained to concentrate more on internal features, there was a noticeable improvement in some cases in their ability to recognize faces. Eye tracking data revealed that after this training, patients began to focus more on the eyes and nose, as well as the area between them, for facial recognition.

1.4 Aim and Research Objectives

The aim of this study is to assess how the process of facial recognition by individuals with Prosopagnosia can inform facial recognition models for accurately detecting similar faces.

- Use medical literature on prosopagnosia to determine relationships between medical assumptions about facial recognition and current deepfake technology.
- Evaluate how deepfake detection techniques are currently implemented and identify any connections to medical literature.
- Examine the coping strategies of individuals with prosopagnosia in managing facial recognition challenges and assess the potential of applying these strategies to improve deepfake detection.
- Perform an experimental analysis by individually occluding different facial features during the training of AI models to assess their impact on deepfake detection accuracy.

- Review existing research on prosopagnosia to identify the specific facial features that individuals with the condition find most challenging to recognize, and analyze how their facial perception differs from that of individuals without the condition.

1.5 Main Contributions

This thesis has explored the application of medical insights to improve deepfake detection methods, resulting in several significant contributions. These contributions are as follows :

- Our experiments, supported by the findings of others, have shown that focusing on facial features beyond the centre of the face can improve the accuracy of deepfake detection (Publication 3).
- We propose Face-Cutout, a novel occlusion technique inspired by strategies derived from the study of prosopagnosia. Face-Cutout leverages facial landmarks and underlying image data to strategically determine cutout regions, enhancing the performance of deepfake detection models (Publication 3).
- Our investigation provided key insights into the effectiveness of deepfake detection models when different facial regions were occluded. Notably, our findings revealed that the model achieved its highest accuracy when the nose region was obscured, followed by the mouth region. This evidence suggests that the nose and mouth are comparatively less critical for deepfake detection, offering valuable guidance for refining detection strategies and improving model performance (Publication 1).
- We used medical literature on prosopagnosia to determine relationships between medical assumptions regarding facial recognition and current deepfake technology (Publication 2).
- Our analysis suggests that the operational mechanism of the AI model differs from the cognitive processes of individuals with prosopagnosia. Specifically, the facial cues that significantly improve face recognition in prosopagnosia patients appear to be less informative to the AI model compared to other facial cues. This distinction underscores the complex nature of face recognition in both humans and artificial intelligence. However, in the context of prosopagnosia, both AI

algorithms and medical approaches concur on the critical importance of the eye region for facial recognition and analysis (Publication 2, 1).

- Our approach demonstrates potential in mitigating the common overfitting problem frequently encountered in deepfake datasets. By introducing variability through diverse facial cutouts at different facial locations, this augmentation method generates distinct iterations of the original image, effectively addressing overfitting issues (Publication 1).
- The research demonstrates that both the EfficientNet B7 and Xception models achieve higher accuracy in detecting deepfake images when utilising the Celeb-DF dataset, compared to their performance with the FaceForensics++ (FF++) dataset (Publication 2, 1).

1.6 Publication

1.6.1 Conference Papers

1. Alanazi, F. (2022). Comparative Analysis of Deepfake Detection Techniques. Presented at the 14th International Conference on Computational Intelligence and Communication Networks (CICN) (pp. 119-124). IEEE.
2. Alanazi, F. (2023). Strategies for Addressing Prosopagnosia as a Potential Solution to Facial Deepfake Detection. Presented at the 12th International Conference on Digital Image Processing and Vision (ICDIPV 2023).
3. Alanazi, F. (2024). Comparative Analysis of Internal and External Facial Features for Enhanced Deepfake Detection. Presented at the 16th International Conference on Agents and Artificial Intelligence (ICAART 2024).

1.6.2 Journal Papers

1. Alanazi, F., Ushaw, G., & Morgan, G. (2023). Improving Detection of DeepFakes through Facial Region Analysis in Images. Electronics, 13(1), 126.

1.7 Thesis Structures

This thesis comprises a total of five chapters. Chapter 1 serves as an introduction, setting the stage for the investigation. It presents the subject matter of deepfake detection, offering an insightful overview of the technology and its implications. This chapter outlines the research objectives, framing them within the broader context of this doctoral investigation, and articulates the problem statement, laying the groundwork for the subsequent chapters.

Chapter 2 delves into the exploration of the intricate relationship between deepfake technology and prosopagnosia. This chapter provides a comprehensive examination of the commonalities shared by these domains, with a particular emphasis on the facial differentiation techniques employed in face recognition. It aims to establish a foundational understanding of how deepfake technology can intersect and interact with the cognitive aspects of face recognition disorders.

Chapter 3 outlines the comprehensive methodology employed in this study, which investigates the intersection of prosopagnosia and deepfake detection. The research seeks to assess whether medical insights into face recognition can enhance deepfake detection techniques by pinpointing critical facial features that distinguish authentic faces from deepfakes. The methodology is structured into a seven-stage process encompassing a literature review, technology assessment, practical experimentation, and data analysis. This approach leverages renowned datasets, including FaceForensics++ and Celeb-DF, and incorporates state-of-the-art deep learning models such as EfficientNet-B7 and XceptionNet. Additionally, a specialized face cut-out technique is introduced to emphasize the most informative facial features, offering a precise and adaptable framework that contributes to future advancements in deepfake detection research.

Chapter 4 presents a comprehensive analysis and critical evaluation of the experimental findings from this study. It systematically details the results obtained across the three distinct phases of experimentation, offering an in-depth examination of the data. Additionally, this chapter provides a comparative evaluation of the proposed methodology against established deepfake detection techniques, highlighting both the strengths and limitations of the approach. By doing so, it offers a nuanced understanding of the efficacy of the proposed methods and their potential contributions to the field.

Finally, Chapter 5 offers a detailed summary of the pivotal findings derived from this research. It delves into the broader implications of these findings, providing a

critical analysis of how they contribute to the existing body of knowledge in the fields of deepfake detection and prosopagnosia. Additionally, this chapter acknowledges the inherent limitations of the study, offering a transparent evaluation of areas where the research may have constraints or where the findings should be interpreted with caution. Building on this, the chapter also proposes potential directions for future research, highlighting areas that could benefit from further exploration and study to continue advancing our understanding of these complex topics.

Chapter 2

Literature Review and Related Works

2.1 Introduction

Advances in deepfake technology have introduced significant challenges in the realms of security and identity verification. As the sophistication and realism of computer-generated faces continue to advance, the potential for using fake facial images in deceptive practices has grown considerably. This evolution makes it increasingly difficult to distinguish authentic imagery from fabricated ones, thereby raising serious concerns in various sectors, including cybersecurity, media integrity, and personal privacy. In response to these emerging threats, it is imperative that the technology dedicated to the detection and identification of deepfakes evolves in parallel, enhancing its capabilities to effectively counteract the growing sophistication of these digital deceptions.

Individuals with prosopagnosia face significant challenges in recognizing faces, including those of familiar people [9]. This impairment can severely impact social interactions and emotional connections, as facial recognition is crucial in human communication. Recent research has provided valuable insights into this condition. Studies have shown that the impact of prosopagnosia can be somewhat alleviated by identifying specific parts of the face and particular facial movements that are more likely to trigger recognition. By focusing on these key features, tailored coping strategies can be developed. These strategies can help individuals with prosopagnosia better navigate social environments and improve their ability to recognize others.

This approach not only offers a practical means of managing the condition but also contributes to a deeper understanding of the cognitive processes involved in facial recognition. In this chapter, it is suggested that findings from studies on mitigating the effects of prosopagnosia could be applied to the detection of deepfake faces. Research on prosopagnosia mitigation is reviewed with the aim of identifying the facial features and movements that are most effective for face recognition. This work is considered within the context of applying these coping mechanisms to the field of deepfake facial detection. The hypothesis proposed is that the facial features and movements that are most useful for facial recognition in individuals with prosopagnosia may also be the most effective for distinguishing a deepfake face from a real one.

2.1.1 The Influence of Media Accessibility on Deepfake Creation

The recent explosive growth in the use of cost-effective intelligent devices, including digital cameras, laptops, tablets, and cell phones, has led to a massive increase in the production of digital multimedia content, such as audio, images, and videos [10]. These smart devices are equipped with operating systems that support applications capable of modifying multimedia content. Acknowledging the impact of these devices and the surge in content generated through these applications is crucial. This proliferation contributes to the prevalence of the post-truth era, where truth is increasingly displaced by alternative narratives facilitated by technology [11].

Digital services provide platforms that enable users to create, share, and distribute digital assets, which may include combinations of text, images, sound, and videos. Due to the lack of strict regulations on the reproduction of digital assets, they are often widely replicated and distributed with minimal user expertise. As digital assets are shared extensively, their provenance can become questionable, complicating efforts to verify their authenticity. Although questionable provenance does not necessarily indicate that an asset is fake, it fosters an environment where content can be easily altered and manipulated without detection. This ambiguity makes it increasingly difficult to determine whether the content is genuine, thus raising the potential for the distribution of fake assets [12].

Provenance is a general term that indicates the perceived past ownership of an item. For a digital asset, provenance is generally more difficult to ascertain. For example, a movie distributed by a publisher using Digital Rights Management (DRM) may not be

considered fake, whereas versions of the movie not protected by DRM are likely to be fake [13].

The inability of individuals to recognise digital fakes and the malicious intent behind such fakes can have detrimental social and economic consequences. For example, DRM removal may hinder revenue streams for a movie publisher, or altered or fabricated content may portray scenarios that mislead viewers into believing falsehoods. If viewers are particularly sensitive to such falsehoods, this could lead to significant social upheaval and inappropriate reactions [13].

One area of fake content that has witnessed a significant rise in popularity over the past decade is the concept termed 'deepfake.' Deepfake refers to the creation of content through the aid of machine learning that appears to represent reality [1].

The term 'deep' in deepfakes is derived from the concept of deep learning, a subset of artificial intelligence techniques pivotal in creating these digital simulations, as noted by Masood in his study on deepfakes [10]. While deepfake technology can be applied to various digital assets, it is most commonly associated with videos and their accompanying soundtracks. At their core, deepfakes are sophisticated, falsified animations crafted using advanced deep learning algorithms to closely imitate real-life scenarios.

In parallel, the rise of social media platforms has significantly contributed to the ease of capturing and sharing digital multimedia content. This ease of content creation and distribution has been a key factor in the proliferation of deepfakes. The rapid sharing of videos, audio, and images across these platforms has made it simpler for deepfakes to circulate and potentially mislead viewers. This phenomenon is further emphasized by Dagar and Vishwakarma [14], who highlight the vast amount of digital content, including deepfakes, that inundates the online space. The intersection of deep learning technology and social media has thus created fertile ground for the spread of these digitally altered realities, posing new challenges in distinguishing between authentic and fabricated content.

2.1.2 The Role of Machine Learning Usage on Facilitating Deepfake Creation

The field of machine learning (ML) has experienced significant advancements, especially with the development and integration of complex algorithms such as generative

adversarial networks (GANs). These advanced algorithms have the capability to manipulate multimedia content with ease, leading to the widespread dissemination of pseudo-information across various social media platforms. The ease with which these algorithms can alter images, videos, and audio has opened the door for their misuse, particularly by those with malicious intent [1].

Individuals with malevolent objectives often exploit these machine learning techniques to tamper with information, aiming to manipulate public opinion or distort reality. This manipulation can take various forms, ranging from the defamation of public figures to political interference, and even inciting public unrest. The use of ML in such contexts is particularly concerning due to its ability to create highly convincing and yet entirely fabricated content. As Akinosho et al. point out in their study [15], the implications of this technology are profound, affecting not just individual reputations but also the broader socio-political landscape.

Moreover, the increasing accessibility and sophistication of these ML algorithms mean that the creation of deepfakes is no longer limited to experts. This democratization of technology has led to a surge in the production and circulation of deepfakes, making it increasingly challenging to maintain the integrity of information online. As a result, there is a growing need for the development of countermeasures, including more advanced detection methods and legal frameworks, to combat the spread of disinformation and protect the public from the potentially harmful effects of these manipulated digital contents [16].

The ease with which media content can be manipulated has made it difficult to trust social media content and, most significantly, identify what is true. The most common standard of proof in court is multimedia content, which is recognized in every legal sector. Therefore, this manipulation places the legal sector in a dire and challenging situation due to the prevalence of media manipulation.As a result, it is critical that all audiovisual content presented as evidence in every legal system is thoroughly examined and verified in order to make sure that it has credibility and integrity [16].

The ease with which media content can be manipulated has made it difficult to trust social media content and, most significantly, identify what is true. Multimedia content, recognized in every legal sector, is increasingly used as a standard of proof in court [17]. Therefore, this manipulation places the legal sector in a dire and challenging situation due to the “potential for media manipulation. As a result, it is critical that all audiovisual content presented as evidence in every legal system is thoroughly examined and verified in order to ensure its credibility and integrity [16].

Furthermore, despite the challenges associated with the authentication of evidence, there is also an emerging technology which is gaining popularity in the field of Artificial Intelligence (AI). This is known as Deepfakes, and it entails the alteration of audio and visual contents through the application of AI-based synthetic processes [18]. Similarly, several other manipulation tools have emerged, such as Sound Forge [19], FaceApp [20], REFACE [21], and Audacity [22], among others, which have increased the difficulty in identifying original digital content. The manipulative possibilities deepfakes provide make them a regular tool for misinformation, consequently making it difficult for the ordinary person to distinguish between fake videos and the originals.

2.1.3 Academic Dissemination of Deepfakes

The recent advances in deep learning and its algorithms have made the subject of deepfakes one of the hottest topics in the field of technology, and have engendered several research projects in recent times [14]. The number of papers in the area of deepfake research, according to a year-wise publication count, and the number of publications by year belonging to the categories studied, have been obtained using Google Scholar. According to the profiling of research publications, interest in deepfakes has grown significantly over the past six years, starting in 2018, as shown in Figure 2.1. This figure provides a visual representation of the increasing number of publications related to deepfakes, highlighting a significant growth in academic and industry focus on this topic. The data, as depicted in the figure, clearly show an upward trajectory in the volume of research, indicating a heightened awareness and concern about deepfake technology within the scientific and technological communities. This increase reflects the growing importance of understanding and addressing the challenges posed by deepfakes in various fields, including digital media, cybersecurity, and information integrity.

2.2 What are Deepfakes?

Various definitions of the concept of deepfake have been proposed by different authors. Emphasising the concept's evolution through different phases highlights the need for a comprehensive examination of its history. According to Dagar and Vishwakarma, the term deepfake 'is a combination of two words, which are "deep" and "fake", connoting

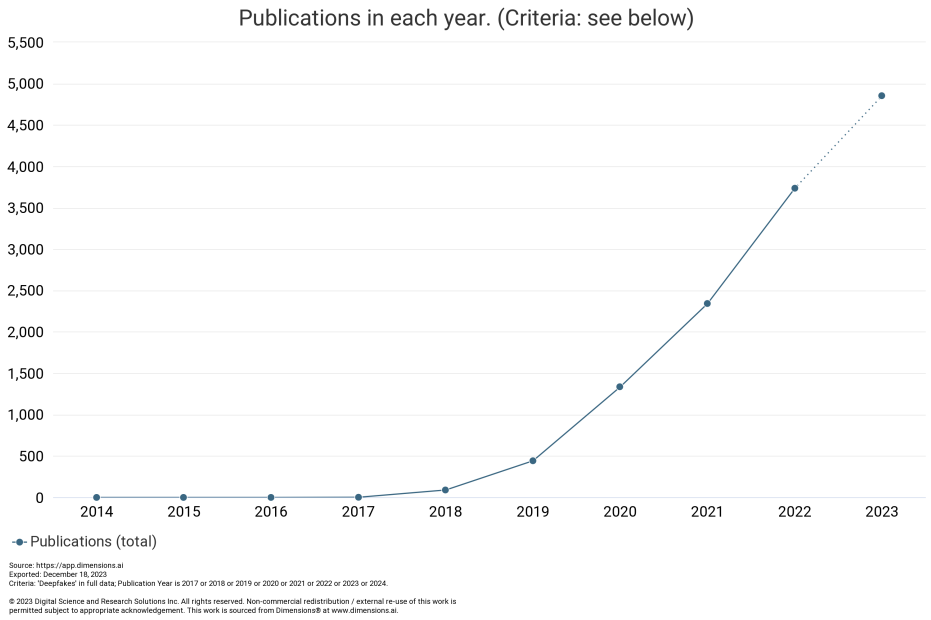


Figure 2.1 Number of Deepfake-Related Publication Articles by Year

fake media that has been modified through the use of algorithms from deep learning, a branch of machine learning' [14].

Nguyen et al. [23] further define deepfake as a term derived from 'deep learning' and 'fake', where a technique is utilised to swap the face of a person targeted in a source video with that of another individual, making the targeted person say and do what the source person said and did. Highlighting global interest among researchers, the creation of deepfakes has notably surged in recent times [10].

These definitions illustrate the manipulative potential of deepfakes, but they focus primarily on videos. Media content created through deep learning techniques is often indistinguishable from original content to the naked eye. Tolosana et al. [1] define deepfakes specifically as videos where one person's face is replaced with another's using deep learning to misinform the public. However, their definition's focus on video deepfakes is narrow and does not encompass the entire scope of the phenomenon, which includes static images and audio manipulations.

While Tolosana et al.'s definition accurately describes the manipulative potential of video deepfakes, it is limited in its scope and fails to consider the broader applications of deepfake technology. Deepfakes also significantly impact static images and audio, where similar techniques create realistic but false representations. For instance, Dagar and Vishwakarma [14] describe deepfakes as 'media that has been manipulated using

deep learning techniques to create realistic alterations. This includes modifying facial attributes, reenacting facial expressions, and generating synthetic audio. Deepfake technology allows for the seamless integration of these changes, making the resulting media difficult to distinguish from authentic recordings.

Similarly, Alanazi and Asif [24] define deepfakes as 'media created using advanced machine learning algorithms to produce highly realistic but fake images, videos, and audio. By training on large datasets, these algorithms can generate media that mimics the appearance and behaviour of real individuals, effectively altering the original content to depict false scenarios.'

These broader definitions highlight the versatility of deepfake technology across various media formats, emphasising the challenges of detecting manipulated content. They suggest that deepfakes are not confined to videos alone but also include images and audio. This broader perspective is crucial for understanding the full concept of deepfakes.

For the purpose of this thesis, which focuses on deepfake images, the term 'deepfakes' will be used in the experimental sections to refer to synthetic images created using deep learning techniques that convincingly alter facial features and expressions to create false but realistic representations. This definition aligns with the broader understanding of deepfakes, acknowledging their potential across different media while concentrating on the specific challenges and implications associated with static images.

2.2.1 Origin and Historical Development of Deepfakes

The origin of deepfakes can be traced to computer vision technology. Computer vision can be described as a complex field that processes images, aiding the computer system with the tools needed to generate information from pictures. Spanning diverse domains, its utility encompasses healthcare diagnostics, the autonomous vehicle sector, and even facial detection applications, as demonstrated by its incorporation into Facebook's photo tagging suggestions. These features establish the fact that deepfake technology can be categorised as within the field of computer vision [25].

The foundation of deepfakes was laid in 1997, when Bregler et al. [26] established the Video Rewrite Program. This program could generate newly found facial imitations from the output of audio. The paper that described the program is the authentic source that first articulated the processes involved in synthesising realistic deepfakes. Cootes et al. [27] then described their active appearance model (AAM) algorithm, which uses

a robust statistical prototype to suitably align with the shape of an image in order to achieve a slight deviation from a source image. This became a significant contribution to the face-tracking and matching space of images. An important contribution in the field of deepfakes was made by Theis et al. [28] by creating Face2Face, which created realistic and convincing deepfakes without the need for any manual intervention by using a combination of deep learning and computer vision techniques.

The term 'deepfake' itself first emerged in 2017 when a Reddit subscriber who called himself Deepfakes claimed to have developed an algorithm in machine learning which could supplant the face of celebrities with the faces found in pornographic videos. He was later banned by Reddit after he started posting images and videos of celebrities through the use of open-source face-swapping tools. The situation eventually devolved into the use of synthetic media applications that could create faces of real people who did not actually exist. Furthermore, it has triggered the need for more research to uncover the diverse applications of the concept, particularly in the area of detection [29].

2.3 Application of Deepfakes Technology

Deepfakes technology, characterized by its versatility, has the potential for both beneficial and harmful applications. The unethical use of deepfake technology poses significant threats to our society, impacting both the present and the future. Everyday social media users are particularly at risk of being deceived or manipulated by deepfakes. Despite these concerns, there are scenarios where the effective application of deepfake technology can yield considerable benefits. The following sections provide detailed descriptions of both the detrimental and advantageous uses of deepfake technology, offering insights into its dual nature and the diverse implications it holds for society.

2.3.1 Beneficial Applications of Deepfakes

Despite the prevalent concerns surrounding deepfake technology, it also has potential applications that can contribute positively to societal welfare. A notable example of this is the Malaria Must Die Initiative, which leveraged deepfake technology to great effect. In this campaign, deepfake technology enabled David Beckham, a prominent public figure, to appear in promotional videos speaking nine different languages [30]. This innovative use of deepfakes significantly enhanced the reach and impact of the

initiative, demonstrating how the technology can be harnessed for humanitarian and global health causes.

Deepfake technology also benefits those with speech impediments due to problems with their speech organs, by developing software such as ALS that can regenerate their voices artificially [5]. Similarly, such techniques could potentially be used for those who are bereaved, by virtually conversing with loved ones through this technology even though they have passed away [31]. In the area of cinematography, deepfake technology has proven to be a significant innovation, especially for portraying deceased actors or de-aging living actors for flashback scenes. This technology can help movie producers recreate younger versions of actors or bring back actors who have passed away, creating realistic and seamless integrations into new films [32].

In addition to the previously mentioned benefits of deepfake technology, museums may utilise this technology to enhance the appeal of their exhibits to visitors. Additionally, history lessons can be brought to life by using images of historical figures to reinforce educational training presentations [5]. The applications of deepfake technology are explained in more detail in the following sections.

Education

Educators can utilise deepfake technology to seamlessly impart requisite knowledge to their students. For instance, historical personalities for whom existing videos are of low quality or unavailable, such as Nelson Mandela or Mahatma Gandhi, can be made available in the form of their teachings about their works [14]. In 2018, a video was produced with Barack Obama warning about deepfakes. This video was direct and appropriate for educating the public about deepfakes [5]. The utilisation of such technology in education could make content more compelling for students.

Entertainment

The entertainment sector has also been greatly influenced by the application of deepfake technology. Other languages can be easily introduced into movies, animating cartoon characters or deceased actors, memes, and the implementation of unique effects in movies [33]. Given the degree to which deepfake technology is evolving, the movie industry will experience a higher level of its application in the future.

Expression

People with speech impediments, such as Amyotrophic Lateral Sclerosis (ALS), can leverage deepfake technology to enhance their communication capabilities through videos enabled by deep learning algorithms.

This technology can also be used to create avatars that allow people to experience virtual worlds that would be impossible to experience physically, such as in video games [5]. It also has applications in communication, such as when a speaker's dialect may differ from that of the audience, and an algorithm could translate the speaker's language into the different languages of the audience for more effective communication [34].

Innovation

Organisations have leveraged deepfake technology's possibilities to attract targeted customers to their brands. For example, deepfakes, through the use of AI algorithms, have created the possibility for customers in the fashion retail industry to virtually turn themselves into models to check and try on new clothes. Furthermore, Reuters has utilised an AI-virtual presenter to broadcast sports news. Similarly, a Japanese firm named Data Grid has started to use a virtual model simulated by AI for its advertising initiatives [34]. The future will be filled with diverse innovative initiatives using the possibilities deepfake technology offers, especially in branding and advertising.

Despite such benefits, the negative effects of deepfakes are expected to considerably outweigh their positive features. Deepfakes are considered to be one of the most serious criminal threats that have evolved through the use of AI. This is because they can be used to create highly realistic and convincing videos that can be used to deceive people [14]. The malicious application of this technology is further elucidated next.

2.3.2 Malicious Applications of Deepfakes

The core aspect of this technology lies in the potential for its misapplication and misuse. Vasist and Krishnan [11] identify three categories in which deepfakes can be harmful, which include harm to subjects, viewers, and institutions. Citron and Chesney [5] further provide a classification of impacts at the individual, organisational, and societal levels. It is important to note that individual and organisational-level harms essentially

entail blackmail, humiliation, instigation, or sabotage through reputational damage. The malicious use of deepfake technology and its impact at the societal level include the manipulation of electoral results, disruption of the democratic process, deflation of public security, and deflation of journalism as a core profession in national security and safety.

Threat to Individuals

Deepfakes hold huge potential to inflict substantial harm, physical discomfort, and psychological stress on their victims. Malicious users can leverage the technology to extract valuable information that could lead to harm. Such malicious users place their victims under unnecessary stress and demand money, business secrets, and personal bank details to prevent them from publishing such information in the public space [5].

Deepfakes can be used to create pornographic videos that exploit people's images and likeness. Moreover, deepfakes may be applied in the workplace to damage the reputation of individuals by depicting them as engaging in antisocial behaviour, such as making racist remarks or abusing co-workers. The future aspirations of individuals can be damaged through the release of deepfake videos purporting to represent evidence of sexual abuse or harassment.

Threat to Business

One of the possibilities deepfakes offer is the ability to swap voices as a form of impersonation, such as the simulation of the voices of CEOs and business leaders to carry out fraudulent activities. An example of this occurred recently in the UK when the simulated voice of a CEO was used to instruct the release of $243,000 to a supplier [35]. Deepfakes can also be used to create imbalances in the market through fake media, enabling business entities to lose or make considerable gains [34]. Furthermore, deepfakes can be used to damage the brand value or product line of an organisation through malicious advertisements, thereby creating an enormous threat to the organisation [35]. Malicious firms can utilise deepfake technology to affect the positions of their close competitors by harming their reputations.

Threat to Society

The central theme of the effect of deepfake technology is its capacity to undermine societal trust. Furthermore, one of its most catastrophic effects is its impact on journalism. The proliferation of the use of social media, coupled with the increase in digital media content and the propagation of deepfakes posted on these platforms, have the potential to create a crisis for society. The higher the prevalence of deepfakes, the more trust in the institutions that disseminate news will erode. The effect of deepfake videos cannot be entirely reversed, even when the truth is discovered. Current evidence shows that deepfakes possess the potential to create panic in society through misinformation, and in extreme cases, can trigger civil war [5].

Threat to Nations

In the context of international relations, bilateral ties can be greatly affected by deepfakes, and the impact may last for generations. Within the realm of international relations, the influence of deepfakes on bilateral relationships can have enduring consequences, potentially resonating across generations. Notably, the utilisation of deepfakes has the capacity to encroach upon both national and international ties. The use of deepfakes by external entities can undermine the democratic process of a nation, leading to civil unrest and the weakening of a nation's security architecture. The release of doctored data through deepfakes can dilute debate on policies, weaken the credibility of speakers, and make it difficult to distinguish between fact and fiction [36].

2.4 Deepfakes Techniques

2.4.1 Deepfakes Creation

The popularity of deepfakes has surged in recent times, largely due to the ease with which they can be created and the proliferation of high-quality doctored videos online. These features are accessible to both professional and novice computer users [23]. The mode of generation of diverse media has influenced the creation mechanism of deepfakes, leading to their classification into various types [37].

The pioneering attempt at deepfake creation was marked by the development of FakeApp by a Reddit user, employing an autoencoder-decoder architecture [38] [39].

In this methodology, Autoencoders involves significant data compression carried out by the network as shown in figure 2.2. Autoencoders can be further divided into three parts:

- Encoder: the function of this part is to extract the features of the input image. It compresses the quality of the input image, most often from a thousandth pixel to a hundredth pixel. Facial measurement is the core function of the encoder, which entails head pose, eye movement, emotional expression, skin tone, and other features[40].
- Latent Space: this exhibits unmatched facial characteristics from which the source image is evaluated and trained. This function focuses more on important facial characteristics. It does not focus on the less important parts of the face, which indicates the image as a compressed kind of the source image, facilitating the memorisation of the important parts of the image [40].
- Decoder: this facilitates the reconstruction of the new image by decompressing the data generated in the latent space to what is very similar to the input image. The performance of the autoencoder is established by the outcome of the comparison between the input and output images and their similarity [40].

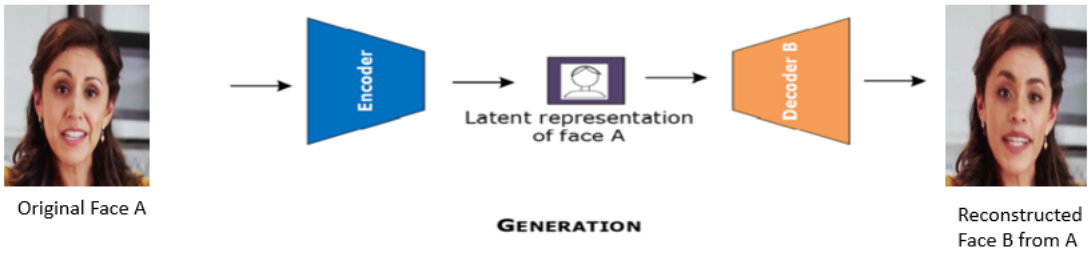


Figure 2.2 Process of Deepfake Generation Using an Auto-Encoder and Decoder

The initial foray into deepfake creation was marked by the development of FakeApp, which used an autoencoder-decoder structure. This approach, exemplified in projects like DeepFaceLab and DFaker, has laid the groundwork for subsequent advancements in deepfake technology [39].

The core of deepfake generation lies in Generative Adversarial Networks (GANs) [41]. A conventional GAN model comprises two neural networks: a generator and a discriminator. These networks are engaged in a minimax game, where the generator aims to produce realistic images while the discriminator attempts to distinguish real images from fake ones. This adversarial training process allows both networks to improve their capabilities over time [23].

Noteworthy among deepfake tools is StyleGAN, introduced by Karras et al. [42], which utilizes a unique generator network architecture for the creation of realistic face images. Unlike traditional GAN models, StyleGAN incorporates a mapping network and a synthesis network, enabling control over image synthesis by modifying styles at different scales. This architecture facilitates the separation of high-level attributes, such as pose and identity, during image generation, offering intuitive control over face synthesis [41].

An enhanced version of deepfakes, known as faceswap-GAN, integrates adversarial and perceptual losses into the encoder-decoder architecture, improving the realism and consistency of eye movements and refining segmentation masks. By leveraging VGGFace for perceptual loss and CycleGAN for generative network implementation, this model enables the creation of outputs with varying resolutions [43].

In essence, the foundation for crafting deepfakes lies in the consistency achieved when the encoder is communicated across two distinct networks. A fake image is created when the compressed version of an input image in the latent space is reconstructed by the decoder with the features of another person's image. This process highlights the sophisticated and evolving nature of deepfake creation, utilizing advanced deep learning techniques to manipulate and synthesize realistic images, posing challenges for detection and raising ethical concerns regarding misinformation and media manipulation.

2.4.2 Deepfakes Detection

The degree of difficulty in discriminating real faces from fake ones has inspired research due to the prevalence of fake digital content that is very difficult to identify [1] . Several detection methods have been proposed to distinguish between fake media content and the original, giving the increasingly negative effects fakes are having on individual lives, democracy, and even the security of society. In the earlier detection methods, hand-crafted characteristics stood prominent, and the detection of fakes entails the

extraction of artefacts and inconsistencies associated with the processes used in the creation of fake videos.

More recent methods involve the utilisation of deep learning algorithms to instantly extract, discriminative, hidden, and inconsistent characteristics in order to detect deepfakes. This type of method is often considered to represent a binary classification challenge in which original videos and fake ones are differentiated using classifiers. These detection methods demand a huge database consisting of both fake and real videos to use in training classification frameworks. There are limitations to creating a benchmark for detecting deepfakes, even though there are more fake videos available. To resolve this puzzle, a renowned deepfake dataset was produced by Fortunian Marcel, which consists of 620 videos that apply the open-source code Faceswap-GAN [44], which is based on the generative adversarial network (GAN) model. The publicly available VidTIMIT database was used to generate videos with both low and high quality, which were essentially deepfake videos that accurately imitate facial expressions, movements of the mouth, and the blinking of the eyes. These videos became the sources of data to examine diverse deepfake detection methods based on VGG [45] and Facenet [46]. According to the test results, it was found that known face recognition systems are not suitable for the effective detection of deepfakes. Also, it is significant to note that high error rates were generated when lip-syncing approaches [47] [48] and metrics of image quality were applied with support vector machine (SVM), from the newly developed dataset. This led to the acknowledgement of the need to develop more suitable and elaborate methods to aid in the detection of deepfakes.

2.4.3 The Classification of Facial Manipulation

The classification of the level of facial manipulation can be categorised in four groups. Figure 2.3 outlines the descriptions of face manipulation based on the level of change of digital assets. The four categories are:

- Entire face synthesis: this form of manipulation utilises GAN to create an otherwise non-existent face. This technique can produce astonishing outcomes of high-quality images. These could be useful indiverse domains, especially in the creation of realistic characters for the video game industry [1] .
- Identity swap: the face of a real person in a video can be replaced by the face of another person using this type of manipulation. Deepfake techniques are utilised to swap the face of the target with the source face [1] .

- Expression swap: this type of manipulation is also known as face re-enactment, which utilises GAN architecture to modify the facial expression of the person. Amongst the common GAN architectures, Face2Face is the most commonly used technique [28].
- Attribute manipulation: this form of manipulation is also known as face editing or face retouching. It entails editing a few facial characteristics so as to achieve minor changes without altering the person's identity. This could take the form of changing the skin or hair colour or adding a moustache or glasses. The FaceApp mobile application is one example of an attribute manipulation tool [49].

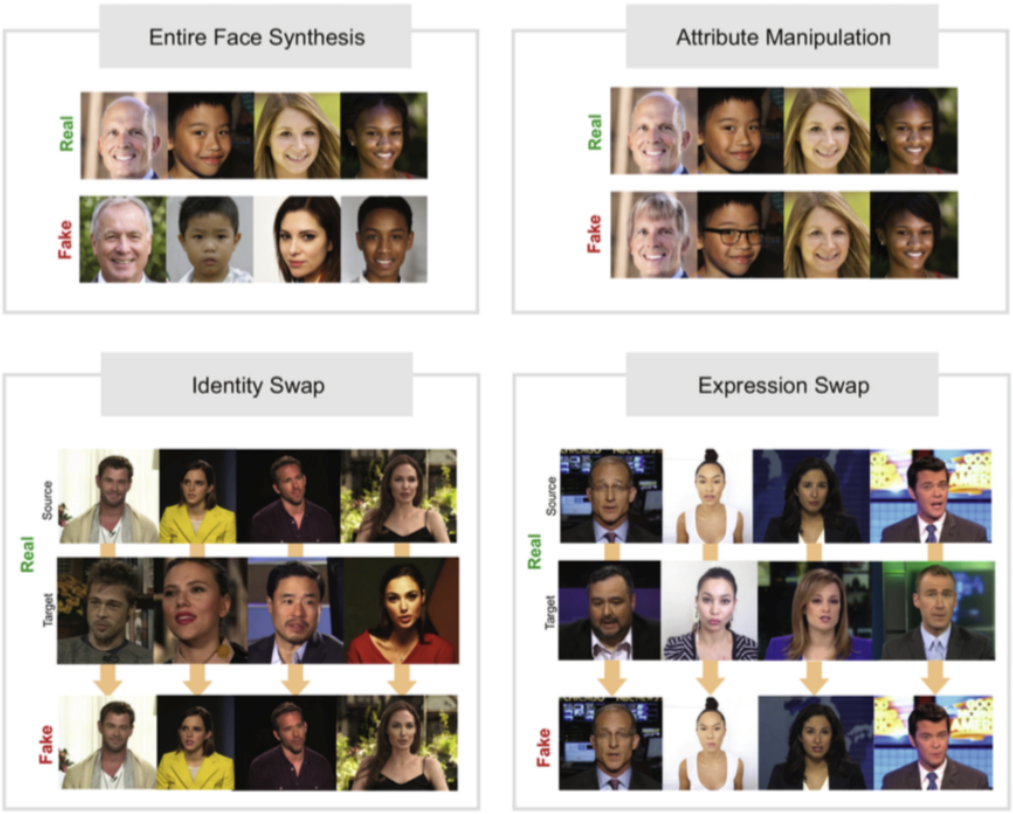


Figure 2.3 Examples of Manipulation Techniques [1]

The focus of this study is the identity swap, in which an in-depth evaluation is conducted of the processes involved in replacing the face of a target person with the face of the source in a video using deepfake techniques which are explored and reviewed.

2.4.4 Detection Clues for the Identification of Deepfakes

The increase in the development of fake images and videos through the utilisation of generative adversarial networks (GANs) has gained prominence recently. According to. Ivanov et al. [4], it is important to acknowledge that deepfakes have evolved over time. In the early phase of their development, it was easy to detect forged images and videos with the naked eye. Since the advent of deep learning technology, it has become challenging to identify images and videos that have been tampered with [50] [51].

The visual artifacts are consistent and represent the elements for detection at the early stage of deepfake evolution. This has also helped in the design of diverse deep learning solutions to detect forged images and videos through the development of algorithms, by identifying the unique artifacts or clues that could be used for detection [51]. This section aims to identify the methods researchers have utilised to detect deepfakes and the detection clues deployed in recent research. A comparative analysis of available deepfake detection methods based on four criteria is shown in Table 2.1, which includes publications between 2015 and 2023.

Digital Anomalies Clues

Deepfake detection primarily hinges on identifying artifacts introduced during the creation process. Various methods, such as convolutional neural networks, are utilized for this purpose. The following studies showcase examples of these anomaly clues used to detect deepfakes.

Detection of Deepfake Images

Hsu et al. [52] developed the Deep Forgery Discriminator (DeepFD), using contrastive loss to identify synthesized images from various GANs, achieving a 94.7% detection rate. Tariq et al. [51] demonstrated that machine learning algorithms, particularly neural network-based classifiers, are effective in identifying fake human faces created by both humans and machines. They used pre-processing techniques and ensemble methods for detecting GAN-forged videos and images.

Techniques	Medium	Detection clues	Year	Accuracy
Dataset: CelebA/CelebDF				
CNN [51]	Image	Detection of modified face regions	2018	94% to 74.9%
DNN [52]	Image	Discriminative features in images generated by the GAN	2018	94.7%
DFT-MF model [53]	Video	Mouth and teeth	2020	71.25%
CNN and SVM [54]	Video	Biological signal map	2020	96%
FD2Foremer [55]	Image	Facial geometry details for deepfake detection	2022	83.81%
CNN [56]	Image	Independent clues as color mismatches, boundary artifacts, and varying quality within images	2023	—
Dataset: FaceForensics++				
RNN [57]	Videos	Temporal discrepancies across frames caused by manipulation of the face	2019	98%
LSTM model [58]	Video	Frame sequence	2020	82%
GANs model [59]	Video	Eye blinking	2020	87.5%
CNN and SVM [54]	Video	Biological signal map	2020	96%
TD-3DCNN [60]	Video	Analyze video frames for inconsistencies	2021	81.08%
Dataset: UADFV				
SVM [61]	Video, Image	Deviation of original face landmarks in the deepfake	2019	—
CNN [4]	Video	Face and head landmarks; Pose estimator	2020	95.5%
Dataset: Various Movies				

Continued on next page

Table 2.1 –*Continued from previous page*

Techniques	Medium	Detection clues	Year	Accuracy
CPBD metric [50]	Video	Sharpness of figure	2015	78.57% to 71.43%
Dataset: Biosec and Warsaw Benchmarks				
CNN [62]	Image	Iris-face fingerprint	2015	98.93%
Dataset: Created by the Authors				
SWIR [63]	Image	Skin detection	2016	49.0% to 73.5%
Dataset: 3DMAD				
PPG and LBP features [64]	Video	Detection of the pulse in the face	2016	86.50% to 95.08%
Dataset: NICT-3D from MERL				
CNN [65]	Video	Track the movement of the pixels	2016	92.36%
Dataset: FERC-2013 and Cohn Kanade (CK+)				
CNN [66]	Image	Observation of facial muscles	2017	—
Dataset: CASIA-FASD and MSU MFSD				
Control the LED light intensity [67]	Video, Image	Reflection of light on the face	2017	69% to 77%
Dataset: Created by the Authors				
LRCN [68]	Video	Detection of eye blinking	2018	—
Dataset: Media Forensics Challenge (MFC)				
SVM [69]	Video	Analysis of multimedia stream descriptors	2019	—
Dataset: CELEBA				

Continued on next page

Table 2.1 –*Continued from previous page*

Techniques	Medium	Detection clues	Year	Accuracy
EM algorithm [70]	Image	Tracing of pixels	2020	88.40% to 99.81%
Dataset: DFDC				
CNN+RNN [71]	Video	Automatically weight different face regions	2020	91.88%
Dataset: IEMOCAP and CMU-MOSEI				
DNN [72]	Video	Detection of sensory noise in face, text, and speech cues	2020	82.7% to 89.0%

Table 2.1 Comparison of DeepFake detection methods, categorized by key criteria such as techniques, media type, detection clues, accuracy, and year of publication. Methods utilizing the same datasets are grouped together, with horizontal dividers for clear differentiation.

Yang et al. [61] focused on the facial region and head poses as regions of interest (ROI) for detection. They noted that deepfakes often involve synthesizing the face region within an original image. Their method uses SVM classifiers to estimate 3D head poses from face images, although it faces challenges in blurred images.

Guarnera et al. [70] proposed the Expectation Maximisation (EM) algorithm for detection, analyzing human faces against GAN architectures. This method discriminates between deepfake and real images by comparing convolutional layers from CELEBA and various GANs.

Li et al. [55] introduced a transformer architecture for forgery detection, beginning with 3D face reconstruction to capture subtle artifacts. This method utilizes a 'MetaFormer' structure and face displacement maps to enhance detection capabilities.

Detection of Deepfake Videos

Kakaletsis and Nikolaidis [50] proposed a technique that utilizes sharpness estimation metrics, extending the Cumulative Probability of Blur Detection (CPBD) metric. This algorithm focuses on the detection of stripes around a human figure's foreground, serving as a key indicator for identifying deepfakes. Their method involves analyzing human

figures in 3D videos by comparing the left and right frames. A crucial detection clue is inpainting, or hole filling, which helps distinguish forged videos from their authentic counterparts using support vector machines and threshold-based classification.

Rana et al. [65] introduced a novel approach for distinguishing between fake and real 3D videos. They highlighted that converting 2D to 3D videos paves the way for deepfake production. Their method uses a convolutional neural network with a dual-tree complex wavelet transform for pre-filtration, generating edge and parallax features to differentiate between fake and real videos. The results showed high accuracy in detection.

Amerini et al. [58] proposed a spatial-based method using end-to-end convolutional neural network classifiers and a sequence-based approach with an LSTM model. This method analyzes consecutive frames to detect correlations, showing promising performance in detecting fake content.

Guerra et al. [69] presented a method combining random forests and SVMs trained on multimedia stream characteristics from both fake and authentic videos. Sabir et al. [57] suggested a two-step process involving face cropping and alignment from video frames, followed by applying a recurrent neural network algorithm to the pre-processed facial region. Their method, tested on the FaceForensic++ dataset, showed a 4.55% increase in accuracy compared to previous models.

Montserrat et al. [71] introduced a technique using a multi-task convolutional neural network (MTCNN) for feature computation and face identification in video frames. The method involves discarding incorrectly detected faces using a gated recurrent unit and an Automatic Face Weighting (AFW) algorithm, followed by an RNN to aggregate features for locating altered information.

In conclusion, while these methods have shown significant success in detecting deepfakes, the rapid advancement in deepfake creation techniques poses a challenge to the long-term sustainability of these detection methods.

2.4.5 Biometric Clues

A variety of human traits are employed in the detection and recognition of fake videos. These include aspects such as facial expressions, eye movements, skin texture, and other physiological characteristics, which are often inaccurately replicated in deepfakes.

Detection of Deepfake Images

Research conducted by Menotti et al. [62] focuses on the application of biometric systems to provide clues for detection through a combination of two approaches: learning using a suitable convolutional network architecture, and reviewing network weights via back-propagation. Several types of spoofing attacks have been developed, which have been deployed for malicious purposes, sometimes leading to situations further degenerating into social unrest. The identification and authentication of people have been aided through the application and development of biometric systems as utilized by international, national, and personal entities as an integral security mechanism. Three important modalities are suggested for the investigation and detection of spoofing: the iris, fingerprint, and face. These detection techniques are based on two algorithms which are used to achieve architecture and filter optimization.

This forms the central focus of research to provide a cross-modal approach which could improve the solutions available for utilization in facial recognition (FR) [63]. The research also employed multispectral short-wave infrared (SWIR) imaging to confirm the validity of faces under investigation, so as to prevent errors caused by partial disguises or facial masks. In establishing the effectiveness of a method, the availability of printers, scanners, makeup, and paints suitable for use in masking makes the forgery process difficult to detect through the application of the visual (VIS) light spectrum. A support vector machine (SVM) classifier using multispectral SWIR imaging helped in accurately detecting skin, where the features of contrast and the brightness of the skin are clues in the detection of forged images. The outcomes of the approach showed a high level of effectiveness in detecting masked faces by drastically reducing the false acceptance rate.

Another area of study focuses on distinguishing fake from real emotions [66]. However, people often hide their true emotions with a sandwiched emotion, and this has been a less developed field of research. The possibility has been explored of distinguishing between emotions through the utilization of a convolutional neural network [73]. Two algorithms were developed to train and test datasets to recognize seven different emotions: happy, sad, disgusted, angry, fearful, surprised, and neutral. Through the utilization of the FERC-2013 dataset, the system was able to distinguish fake smiles from real smiles because the features of fake smiles were seen to be elaborately different from those of real smiles.

A transition from a password approach of control to access to a facial recognition (FR) approach as a suitable alternative has gained prominence in the research community. Research by Mhou et al. [67] identified 3D masks, printed photographs, and video replay attacks to be major problems affecting the utilization of facial recognition, and suggested a robust approach to detect spoofing attacks. The reflectiveness of light rays on different surfaces was the basis of the approach developed through the utilization of Laplacian blur detection, Gabor filters, and local binary patterns. These algorithms calculate the reflectiveness of the light rays on different surfaces and are used to classify real and fake images. Their results show a significant improvement in the detection of spoofing attacks.

Detection of Deepfake Videos

The widespread use of GANs, as established by Hsu et al. [52] and Li et al. [68], further reinforces the malicious use of deep neural networks, which must be countered by a method of detection which assesses eye blinking. It is noted that this feature is a physiological characteristic which is not well-articulated in forged videos. The detection method is based on the long-term recurrent convolutional network (LRCN), with an algorithm developed to monitor the dynamics associated with eye blinking as a clue to detection based on the analysis of video frames. The model shows a high level of performance in detecting videos and images generated by deep learning neural networks. However, it employs solely the absence of eye blinking as a criterion in identifying deepfakes. The following potential drawbacks of this strategy include that forgeries are difficult to spot in videos with frequent eye blinking, or when there are altered faces with closed eyes during training, and in situations where forgers may make synthesized faces blink realistically.

Jafar et al. [53] noted the proliferation of smartphones with digital camera features coupled with apps that can be used to edit and transfer digital content. The advent of artificial intelligence (AI) through deep learning tools now provides functionalities that can be used to compromise and distort the actual characteristics of images and videos for malicious purposes, which has the potential in some circumstances to cause social unrest. Research has suggested a model for the detection of these fake videos and images through the utilization of a convolutional neural network and the DFT-MF technique. The clues for detection include movements of the lips or mouth evaluated by isolating, evaluating and authenticating datasets generated from the Deepfake Forensics (Celeb-DF) and Deepfake Vid-TIMIT datasets. The outcome of the methods and

techniques utilized indicated high-level accuracy in the review of mouth movements when certain words were pronounced. The movements analyzed in fake videos involved a wider and more open mouth in comparison to real videos from the datasets.

Research conducted by Jung et al. [59] focused on using the blinking of the eyes as a clue to identify deepfake videos by deploying an algorithm termed Deep Vision. There are predictable patterns of eye blinking, which is spontaneous and voluntary, and this technique was applied in the investigation of eight videos where detection was successful in seven of them. However, the utilization of combined cues has rarely been explored recently, despite the positive outcomes single cue models have generated.

Ciftci et al. [54] presented a method using biological data to analyze videos and find forensic changes as indicators from the facial area of the videos of variables such as heart rate. SVM and CNN models were trained using the temporal and spatial aspects of facial features to distinguish between false and real videos. Although this method has evolved and improved the precision of deepfake detection, it has a significant drawback: the precision of its detection in video is significantly reduced when dimensionality reduction techniques are used.

Finally, the biometric clues demonstrate promising potential and increased sensitivity in detecting fakes. This is primarily because biometric characteristics, such as facial expressions, eye movements, and skin textures, are inherently complex and unique to each individual. Deepfake technology, despite its advancements, often struggles to replicate these subtleties with complete accuracy. Furthermore, as biometric data is deeply rooted in human physiology and behavior, it provides a robust framework for identifying discrepancies that artificial intelligence-generated fakes typically exhibit. This makes biometric clues not only more reliable but also more adaptable to the evolving techniques used in deepfake creation, offering a forward-looking approach to deepfake detection.

2.4.6 Multi-Clues

These methods employ combinations of various clues concurrently, leveraging the strengths of different detection techniques. These different clues, encompassing both anomalies and biometric indicators, can be used together to enhance the accuracy of deepfake detection. Anomalies might include inconsistencies in lighting, digital artifacts, or unnatural movements, while biometric clues involve more subtle, physiological aspects like facial expressions, eye movements, and skin textures. The following studies provide

examples of how these combined clues can be effectively utilized. By integrating multiple types of clues, the detection process becomes more robust, as it does not rely on a single point of failure. This multifaceted approach allows for a more comprehensive analysis, increasing the likelihood of accurately identifying deepfakes.

Detection of Deepfake Videos

Mittal et al. [72] explored a new approach to the detection of emotions in proposing a Multimodal Emotion Recognition Algorithm (M3ER). It utilizes three cues relating to the face, text, and speech and applies canonical correlational analysis to distinguish between effective and ineffective artifacts. It is noted that one of the challenges associated with the application of this method is the difficulty in identifying cues to combine so as to achieve the desired outcome.

Ivanov et al. [4] conducted a review of deepfake studies in order to identify the approaches utilized in the detection of forged videos and tampered-with images. Significant and consistent improvements were found in the methods and techniques researchers have developed to detect deepfakes, but the need for more elaborate methods and techniques was emphasized. The super-resolution algorithm was proposed as a convolutional neural network algorithm which could expose deepfakes by identifying inconsistent head poses and applying Resnet50 to identify deepfakes. The clues also included eye blinking, mismatched color profiles, and face-warping artifacts as combinations of clues. The outcome of the approach was a 94.1% rate of fake detection for Resnet50, while the estimated head direction vector achieved a rate of 50.1%. The authors proposed an advanced version of the Resnet application for better detection performance.

Mittal et al.'s [72] M3ER method for the detection of emotions cited above utilizes face, text, and speech cues and canonical correlational analysis to distinguish between effective and ineffective artifacts. This approach was tested using the IEMOCAP and CMU-MOSEI datasets, and the results show a mean accuracy of 82.7% and 89.0% respectively.

Another study [60] used multiple techniques to distinguish between real and fake videos. It analyzes physiological signals, detecting inconsistencies like irregular pulse rates across frames, a common shortfall in deepfakes. Additionally, it identifies temporal inconsistencies, focusing on unnatural behaviors within frames, such as abnormal blinking. The advanced method involves Temporal Dropout in 3D Convolutional Neural

Networks (TD-3DCNN), which scrutinizes frames for discrepancies using 3DCNNs enhanced by a temporal dropout feature that randomly samples frames, aiding in the effective detection of deepfakes.

This research [56] introduces a deepfake detection approach focused on recognizing quality discrepancies between patches commonly seen in deepfakes. It aims to improve adaptability by identifying clues that are independent of the domain and effective against different forgery methods. The study detects unique signs in deepfake images, such as color irregularities, noticeable artifacts at synthesis edges, and quality differences between facial and non-facial areas. The method utilizes an interpatch dissimilarity estimator and a multistream convolutional neural network to detect these specific deepfake characteristics.

Overall, the studies effectively demonstrate the strength of this integrated strategy. They highlight how combining anomaly detection with biometric analysis, or employing multiple types of clues, can significantly enhance the reliability and effectiveness of deepfake identification. However, a primary challenge for this approach is the increased complexity and computational demand involved in processing and analyzing various clues simultaneously. This complexity could potentially impact the efficiency and scalability of the detection process.

2.5 Prosopagnosia

This section conducts an analysis of the medical condition known as prosopagnosia, explores methodologies aimed at improving facial recognition capabilities in affected individuals, delineates the facial regions offering pertinent information for face differentiation, elucidates the specific challenges faced by individuals with prosopagnosia in recognizing certain facial features, expounds upon their perceptual experience, and discusses the prospective application of acquired insights to advance techniques for the detection of deepfakes

2.5.1 How Does the Brain Distinguish Faces

Despite the apparent ease and seemingly effortless nature of face recognition, an extensive body of research has delved into unraveling the intricate mechanisms within

the brain responsible for this cognitive function. The process of facial recognition is intricate, constituting a sequential chain of processes with distinct stages [74].

To elaborate, the journey of facial recognition commences with the perceptual analysis of facial characteristics. This initial phase involves the nuanced encoding of unique features that collectively form the individual's facial signature. Following this, the facial representation is internally generated within the brain, presenting itself as a distinctive figure. The subsequent stage involves the comparison of this internally generated facial representation with the stored memory of familiar faces. This step is critical in determining whether there is a match. In the event of a successful match, the brain can effectively extract details about the person from long-term memory. This retrieval encompasses not only facial features but also associated information, contributing to a comprehensive understanding of the individual [75].

In essence, the cognitive process of facial recognition encompasses perceptual analysis, feature encoding, mental representation, and memory retrieval. The orchestration of these processes highlights the brain's remarkable ability to seamlessly and efficiently recognize and differentiate faces, underscoring the complexity and sophistication inherent in human cognitive function [76].

2.5.2 Face Processing Based Neurobiology

The concepts of face recognition and perception are important and integral components of social interaction and refer to significant skills that are acquired in infancy. Facial perception and recognition induce behavioural patterns due to their influence on levels of attraction, familiarity, and emotional status. It is important to state the fact that facial recognition occurs approximately 70 milliseconds after stimulus presentation as a spontaneous and robust process in humans. From the age of two months, newborn babies track faces as a distinct innate processing capability when compared with other stimuli [77].

Facial judgement, which grows alongside face recognition, emerges early and establishes the significance of face processing as a central element of human social life and survival. There is a connection between the structure of the brain and human behavioural patterns based on the concept of face processing. Studies conducted in the 1970s and 1980s show that face processing is connected to different brain networks involved in the recognition of familiar and unfamiliar faces and face discrimination. Subsequent studies in facial recognition and specialization took another dimension

based on the tools and algorithms available in the early 1990s, during which period Charlie Gross demonstrated the existence of networks and regions in the brain that were possibly involved in face processing [78] [79].

One of the profound discoveries made by Gross concerns the presence of cells in the inferior temporal cortex that induce response mechanisms to the face and hands. This confirms the existence of 'grandmother' cells, as speculated earlier, which respond to specific stimuli that are meaningful. The discovery of an area selective for face perception in humans in the fusiform gyrus resulted from the application of functional Magnetic Resonance Imaging (fMRI). Further fMRI studies in macaque monkeys confirmed this finding, and single-cell recordings showed that neurons in this region respond selectively to faces [79].

Further studies in human face processing reveal that individual faces stimulate specific reactions shown as patterns in the anterior momentary cortex. Other researchers have confirmed that an anterior temporal network and the fusiform region are implicated in the facial recognition process, and a central role may be played by the ventral anterior temporal lobes. Meanwhile, a disorder of face processing creates an impaired ability to match faces and make judgments of facial expressions. This medical condition, resulting from the impairment of specific cells in these brain regions, is called prosopagnosia [80].

2.5.3 Causes, Symptoms, and Types of Prosopagnosia

Prosopagnosia, commonly known as face blindness, is a neurological disorder marked by the inability to recognize faces. This condition can be categorized into two main types: acquired and developmental (congenital). Acquired prosopagnosia arises from brain damage, which can result from various causes such as stroke, traumatic brain injury (TBI), neurodegenerative diseases like Alzheimer's, brain surgery, or encephalitis. These incidents lead to damage in critical areas of the brain involved in face processing, such as the fusiform face area (FFA) and the anterior temporal lobes, impairing the individual's ability to recognize familiar faces [81].

Developmental prosopagnosia, on the other hand, is present from birth and does not result from any evident brain damage. The potential causes include genetic factors, as this form of prosopagnosia can run in families, suggesting a hereditary basis. Additionally, abnormal brain development, particularly in regions like the fusiform face area (FFA) and the anterior temporal lobes, can lead to this condition. These brain regions play vital roles in processing and recognizing complex visual information,

including faces. Damage or abnormalities in these areas disrupt the normal face recognition process, leading to the symptoms of prosopagnosia [81].

The Symptoms of , manifests primarily as an inability to recognize familiar faces. People with this condition can see and distinguish facial features (eyes, nose, mouth) but cannot link these features to an identity. This leads to a lack of familiarity with faces they should recognize, making it difficult to retrieve personal information or names associated with these faces. Individuals often rely on non-facial cues such as voice, clothing, hairstyle, or gait to identify people. In severe cases, prosopagnosia can extend to difficulties recognizing one's own face in the mirror [7].

2.5.4 Prosopagnosia and Facial Recognition

Prosopagnosia, also known as 'face blindness,' hinders the ability to recognize familiar faces [82]. Recent studies highlight the challenges faced by individuals with this condition in matching faces and judging facial expressions [83] [84] [85] [86]. An object is first examined before a judgment is made, with the orientation and judgment processes being different. Results show that unusual and normal arrangements of facial patterns are not significantly different. The research further establishes that the subjects had difficulty recognising familiar faces, as well as unusual and normal arrangements of objects [87].

Facial recognition is a complex process influenced by both the left and right hemispheres of the human brain [86]. Structures in the left hemisphere have an impact on the recognition of facial expressions, while, in contrast, structures in the right hemisphere impact recognising components of figure and form that are critical to facial recognition [88].

The brain undertakes a feature-by-feature analysis when a non-verbal component such as a face is the stimulus, and this process takes place in the brain's left hemisphere, whereas familiarity is assessed in the right hemisphere. Sequential presentation of stimuli such as the face is considered more effective in recognition than a typical presentation because it allows the left hemisphere to analyze the features of the face one at a time. This is in contrast to a typical presentation, where the entire face is presented at once, especially for an individual whose right hemisphere is damaged [83]. From evaluations of the ability of people with prosopagnosia to recognise faces and facial expressions, it is evident that sufferers can differentiate between typical and Thatcherized faces and have a partial ability to recognise facial expressions.

(Thatcherization consists of a face image wherein the eyes and mouth have been turned upside down relative to the rest of the face.)

Research results point to the patients having lost their configural processing ability, affecting their ability to categorise typical and Thatcherized faces. However, they had intact feature processing ability that supported the categorisation of facial emotion and differentiation between typical and Thatcherized faces [85]. Prosopagnosia has significantly influenced the trajectory of face recognition research. By challenging conventional models of facial perception, this condition has prompted a more nuanced understanding of the underlying cognitive processes.

However, the study of prosopagnosia has revealed the complexity of facial processing, demonstrating the involvement of multiple neural pathways. As highlighted by Stone and Valentine [89], investigations into prosopagnosia have challenged established models, such as those proposed by Burton et al. [90] and Farah et al. [91].

Furthermore, research on prosopagnosia has been instrumental in pinpointing the neural correlates of face recognition. Neuroimaging studies, as cited by Stone and Valentine [89], have implicated the fusiform face area (FFA) as a critical region for facial processing. The presence of abnormalities in this area among individuals with prosopagnosia underscores its significance in face recognition.

Beyond structural brain differences, prosopagnosia research has shed light on the role of affective factors in facial perception. Studies by Greve and Bauer [92] have demonstrated that individuals with prosopagnosia can exhibit preferences for familiar faces without explicit recognition, suggesting the influence of emotional responses on facial processing.

In conclusion, prosopagnosia has been a catalyst for advancements in face recognition research. By challenging existing paradigms, informing neurobiological investigations, and expanding our understanding of facial perception, this condition has contributed significantly to the field of cognitive neuroscience.

2.5.5 Facial Processing, Eye Movement, and Fixation Patterns in Individuals with Prosopagnosia

An established mechanism for the recognition of a person is to observe their facial features, especially internal components such as the eyes, nose and mouth. According to Henderson, Williams and Falk Henderson et al. [82], people focus more on specific

attributes in the faces of others that give them a unique identity. Van der Geest et al. [93] further confirm that the most compelling features that attract the attention of an individual while looking at the face of another person are the eyes and the mouth. These features are critical in assessing another person's identity, mental state and emotional condition.

However, Schwarzer et al. [94] brought to light a unique approach to face recognition in individuals with prosopagnosia. This condition, characterized by a difficulty in recognizing familiar faces, forces affected individuals to adopt alternative strategies for identification. Unlike typical observers who focus on central facial features like the eyes, nose, and mouth, those with prosopagnosia lean heavily on the external features of the face, such as hair, neck, and chin. This shift in focus is a compensatory mechanism, as central facial features offer limited cues for recognition in these individuals.

This reliance on peripheral features is significant because it underlines a fundamental difference in how people with prosopagnosia process facial information. For the majority, the central features of a face are crucial for identifying and distinguishing one person from another. These features, especially the eyes and the mouth, are rich in detail and are often used to perceive emotional expressions and subtle identity cues. However, for individuals with prosopagnosia, these central features do not provide the necessary information for recognition, possibly due to deficits in the brain areas responsible for processing these complex visual stimuli [95].

Building upon the findings of Bobak et al. [96], it is evident that the nose region plays a crucial role in face recognition, particularly in differentiating individuals with similar facial attributes. This area of the face serves as a pivotal point for holistic and configurational processing, essential for identifying unique facial features. In their research, Bobak et al.[96] found a positive correlation between the time spent focusing on the nose and the ability to recognize faces accurately. This challenges the conventional emphasis on the eyes and highlights the nose as an equally, if not more, important region for facial recognition.

Expanding on these insights, Bate et al. [97] conducted a detailed eye-movement analysis study. Their research focused on individuals with prosopagnosia, a condition characterized by difficulties in recognizing familiar faces. The study revealed that people with prosopagnosia tend to concentrate their gaze more on the nose region when attempting to identify a face. This pattern of eye movement contrasts with typical face recognition strategies, where the focus might be more evenly distributed across different facial features. This observation underscores the significance of the

nose region in facial recognition processes, especially in individuals with prosopagnosia, where traditional recognition strategies might be less effective.

2.5.6 Clues for Detecting Facial Recognition in Individuals with Prosopagnosia

Prosopagnosia, commonly present from birth, poses lifelong challenges for those affected, significantly impacting their ability to recognize friends, family, or partners. Consequently, individuals with this condition often resort to alternative face recognition methods. According to Schwarzer et al. [94], features such as the hair, neck and chin are critical elements used by patients to recognise people in their vicinity. It is worth noting that this strategy requires complex analysis; otherwise, it may not be practical, causing individuals with the condition to avoid social interaction or have an overwhelming fear of social situations. For instance, Bate et al. [98] concluded that individuals with prosopagnosia could easily recognise the faces of others by relying on features external to the face.

Bennetts et al. [95] found that observing a face in motion can assist people in the general population to recognise others, and a similar result was obtained for individuals who rely on movement cues as a supplementary strategy for processing faces. However, there is a need for further investigation to examine the strategy in the context of familiar face recognition. The author found that people are better at recognising faces when they are moving, as opposed to when they are static images. This suggests that people use motion cues to help them recognise faces. This finding could be helpful for people with prosopagnosia, as it suggests that they may be able to improve their face recognition skills by observing familiar faces in motion. They can learn and identify patterns in facial transitions where movement is a vital cue in conditions of face recognition impairment.

Patients with prosopagnosia tend to rely more on the shape of the mouth rather than the structure of the nose to recognise people, as noted by Pizzamiglio et al. [86]. Burra, Kerzel, and Ramon [99] affirm that people with prosopagnosia rely on external features for face recognition because they have difficulty processing information based on the eye region. A study by Diaz [100] also found that the participants relied on non-facial cues such as hairstyle, gait and voice, as well as location, to recognise people. Moreover, Caldara et al. [101] concluded that the lower part of the face, and especially the mouth and external contours, can be instrumental when processing familiar faces.

This is in marked contrast to normal observers who use eye information to identify familiar faces. Fine [102] also identified that features such as unusual clothing or a particular facial element, such as type of moustache, could assist in facial recognition. Meanwhile, Amanda et al. [103] presented evidence for effective face recognition based on eyebrows, blemishes, and other distinctive features such as skin tone.

Notably, patients with face recognition challenges often resort to using multiple cues for identification, as shown in Table 2.2. This strategy, particularly crucial for those with conditions like prosopagnosia, involves combining various features such as hairstyle, voice, and distinctive clothing. This multi-cue approach enhances the likelihood of accurate recognition, compensating for difficulties in identifying faces based solely on facial features.

Source	Recognition Clue in Prosopagnosia	Year
[101]	The lower part of the face, including the mouth and the external contours, as normal observers typically do when processing unfamiliar faces.	2005
[94]	External features, such as hair, neck, and chin.	2007
[100]	Non-facial contextual and visual cues to identify individuals such as hairstyle, clothing, gait, voice, and location.	2008
[102]	Extra-facial information such as unusual clothing, characteristics or particular facial features like type of moustache.	2012
[95]	Alternative sources of information such as the body or movement.	2015
[97]	External features.	2015
[86]	Mouth and nose.	2017
[99]	External features; avoid processing information using the eye region.	2017
[103]	Eyebrows, blemishes, distinctive features, skin tone.	2020

Table 2.2 Face Recognition in Prosopagnosia Condition.

2.5.7 Rehabilitation & Training Programs for Prosopagnosia

Prosopagnosia, presents significant challenges for affected individuals, who often cannot form complete and detailed mental representations of faces. This deficit compels them to adopt alternative recognition strategies, such as focusing on discrete facial features (e.g., the shape of the nose, color of the eyes) or relying on non-facial cues (e.g., an individual's voice or clothing). The repercussions of these compensatory tactics extend far beyond mere recognition difficulties, adversely affecting social interactions, employment opportunities, and overall self-esteem [104].

Historically, the prognosis for enhancing facial recognition capabilities in prosopagnosia sufferers was bleak. A notable hypothesis by Coltheart suggested [105] that certain cognitive deficits, including those affecting face recognition, could be permanent following neurological damage. This view was predicated on the belief that face processing is dependent on a specific, localized region of the brain, and that damage to this critical area could result in enduring functional impairments.

Contrary to these earlier assumptions, recent advancements in research, including contributions from Coltheart's own team [106, 107], have offered a more hopeful outlook. Evidence now suggests that individuals with developmental prosopagnosia can experience improvements in their face recognition abilities through carefully designed training and rehabilitation programs. This section aims to critically review the literature on rehabilitation and training for individuals with prosopagnosia, providing a balanced exploration of both the promising outcomes and the limitations of these interventions. By examining the evidence presented in various studies, as summarized in Table 2.3, we will assess the efficacy of these training programs, highlighting cases of both significant advancements and areas where improvements remain elusive.

Pioneering Training for Prosopagnosia,the earliest documented attempt to improve face recognition in prosopagnosia was conducted by Beyn and Knyazeva in 1962 [108]. Their study focused on a 39-year-old patient (C.H.) experiencing severe difficulties recognizing familiar faces, likely due to bilateral damage in the occipital-temporal regions of the brain. The researchers employed a systematic training program that involved,Focused practice on facial features and expressions: C.H. actively engaged in activities that directed attention to specific facial components and their variations. Beyn reported that C.H. exhibited some improvements in recognizing faces in real-world situations following the training program. While limitations exist in the absence of standardized training methods and objective assessment tools, this case study provides

an initial indication that focusing on specific facial features might be a valuable strategy for mitigating face processing deficits in individuals with prosopagnosia [108].

In contrast,A study [109] conducted by Wilson in 1987 reported disheartening outcomes. This research focused on a 27-year-old subject diagnosed with prosopagnosia, accompanied by right temporal-parietal brain damage. The therapeutic approach adopted involved Practice on facial recognition using visual imagery and motor movements to facilitate face recognition. Despite undergoing 11 evaluative tests, the individual demonstrated no notable enhancement in their ability to recognize faces [109].

Ellis and Young [110] undertook a detailed investigation to explore the feasibility of retraining face discrimination abilities in a child with prosopagnosia. The subject of the study was an 8-year-old, referred to as K.D., who had suffered diffuse brain damage due to meningitis. Over the course of 18 months, K.D. participated in a structured training program designed to enhance systematic face discrimination and face-name association skills, incorporating feedback mechanisms. The researchers posited that rigorous and focused practice with a select group of faces within a controlled setting could potentially enhance K.D.'s capabilities in processing faces.

However, the results were not encouraging. Despite repeated training involving familiar and unfamiliar faces, as well as discrimination tasks with varying levels of difficulty, K.D. demonstrated no significant improvement in face recognition or face-name association. The authors acknowledge limitations in the study, including the relatively low daily training intensity (approximately 10 trials per day) and the lack of initial tasks tailored to K.D.'s ability level, which may have contributed to frustration and reduced motivation. Nevertheless, the findings suggest that once damaged, the face processing system exhibits limited potential for remediation, even in young, developing brains [110].

In 2002, Francis et al.[111] reported some degree of improvement following training in a 21-year-old patient (N.E.) diagnosed with prosopagnosia and person-based semantic deficits, attributed to damage in the right temporal lobe, which was possibly bilateral, caused by herpes encephalitis. Utilizing face learning strategies, it was found that encoding techniques that simultaneously targeted semantic impairments and face processing deficits proved to be the most effective. These strategies not only facilitated the recognition of unfamiliar faces but also improved the recognition of faces familiar to the patient. However, despite these positive results, the authors urge caution, noting that N.E.'s basic face perception abilities remained largely intact. Consequently, the

observed improvements might not extend to individuals with acquired prosopagnosia who suffer from more severe perceptual deficits.

Another study by Brunsdon et al. [107], confirms the efficacy of targeted training in childhood AL. It represents a significant step in understanding and improving face recognition abilities in children with developmental prosopagnosia, highlighting that abnormal scan paths for faces might be a common factor in this condition. The findings also suggest that early targeted training focusing on eye movements could positively impact the development of face recognition abilities in children with developmental prosopagnosia.

Another piece of evidence supporting the improvement of training children with prosopagnosia is found in the study by Schmalzl et al. [106], which focused on training familiar face recognition and analyzing visual scan paths in a child with congenital prosopagnosia (CP) showed significant improvements in the child's attention to internal features of faces post-training. Initially, the child (referred to as K.) directed most of their attention to the nose, with lower accuracy levels in recognizing other facial features. However, after the training, there was an increase in the accuracy of recognizing faces, with a notable shift in the attention pattern.

Post-training, K. spent an average of 90.8% of dwell time on internal features, a significant increase from the pre-training focus. The pattern of attention also changed, with the eyes being fixated more than the brow for familiar faces, differing from the pre-training pattern. The largest percentage of dwell time and fixations remained directed towards the nose, but there was a noticeable shift in focus towards other facial features, such as the eyes and brow [106].

The training led to a flawless recognition of front-view photographs of familiar faces and generalization to photographs from different viewpoints one month later. This suggests an improvement in structural encoding, specifically in K.'s ability to encode facial features and their characteristics within the face gestalt. Repeated practice with the same photographs presumably strengthened and facilitated access to representations of familiar faces [106].

Furthermore, Mayer, E., and Rossion, B. [112] study described a rehabilitation strategy where a patient with prosopagnosia (P.S.) was trained to analyze internal features of faces. The focus of the rehabilitation was on enhancing the patient's ability to recognize familiar faces by paying attention to internal traits such as the eyes, nose, and mouth. This approach was chosen because the processing of internal features, especially the eyes, is essential for recognizing familiar faces and increases with face

familiarity. The training lasted four months, with two sessions per week. The patient was initially taught to identify and describe various facial features, such as almond eyes or a turned-up nose. This training aimed to enhance the patient's ability to focus on and process the internal features of faces, which are crucial for facial recognition, particularly in familiar faces. After four months of training, P.S. showed improved recognition of her students' photographs and increased reliance on internal facial features for recognition. Furthermore, she gained the confidence to accompany her students outside the school, indicating tangible improvements applicable to real-world scenarios.

Another application of a training regimen designed to enhance face perception and recognition involved a 48-year-old individual with developmental prosopagnosia (DP), identified as M.Z. In the study conducted by DeGutis et al. [113], M.Z. engaged in extensive training over several months, completing over 20,000 trials. Post-training, M.Z. exhibited notable improvements on standardized tests, such as the Benton Face Perception Test, and also reported practical improvements in daily life face recognition tasks. These improvements remained effective for several months before diminishing.

Contrasting with the positive results of holistic face processing training in individuals with developmental prosopagnosia (DPs), Dalrymple et al. [114] reported an unsuccessful intervention attempt in an adolescent DP, using a training approach similar to that described by Ellis and Young [110]. In this case, DeGutis and colleagues attempted to train 12-year-old T.M. to recognize his mother's face through a "mom/not-mom" identification task, where T.M. had to differentiate his mother from age-matched females, receiving feedback after each trial. Despite participating in 47 training sessions, each approximately 10–15 minutes long over a span of 10 months, T.M. did not exhibit any notable improvements in the task, nor did he report any improvements in his everyday face recognition skills. This outcome contrasts with results from other studies [57, 107] that demonstrated training-induced improvements in the face recognition abilities of young individuals with prosopagnosia, highlighting potential limitations in the effectiveness of face processing enhancement efforts in DPs, even in those with developing brains.

However, these collective findings still provide compelling evidence that the face processing abilities of individuals with developmental prosopagnosia can, to some degree, be improved through targeted interventions.

DeGutis et al. [115] investigated the impact of intensive training on a 46-year-old patient with acquired prosopagnosia (C.C.) resulting from a lesion in the right

Table 2.3 Case Studies Evaluating the Effectiveness of Rehabilitation and Training Programs on Enhancing Face Recognition Abilities in Individuals with Prosopagnosia

Source	Patient Code	Age/Gender	Duration	Improvements
[108]	C.H.	39 years / Male	11 months	✓
[109]	-	27 years / Male	3 weeks	X
[110]	K.D.	08 years / Male	18 months	X
[111]	N.E.	21 years / Female	14 days, 7 sessions	✓
[107]	A.L.	8 years / Male	1 month	✓
[106]	K.	4 years / Female	over a month	✓
[112]	P.S.	52 years / Female	4 months	✓
[113]	M.Z.	48 years / Female	14 months	✓
[114]	T.M.	12 years / Male	Over 10 months	X
[115]	C.C.	46 years / Female	One month, 30 sessions	X
[116]	N = 24	-	15 sessions over 3 weeks	✓
[117]	-	-	11 weeks	✓
[97]	EM	14 years / Female	One year	✓
[118]	R-IOT1	55 years / Male	11 weeks	X
	R-IOT4	60 years / Male	11 weeks	✓
	L-IOT2	60 years / Male	11 weeks	X
	B-IOT2	60 years / Male	11 weeks	✓
	B-ATOT2	23 years / Female	11 weeks	✓
	B-ATOT3	15 years / Male	11 weeks	✓
	R-AT3	41 years / Male	11 weeks	✓
	R-AT5	61 years / Female	11 weeks	✓
	B-AT1	31 years / Male	11 weeks	✓
	B-AT2	48 years / Female	11 weeks	X
	B-IOT3	48 years / Female	11 weeks	✓

occipitotemporal area. C.C. participated in a rigorous one-month training program consisting of 30 sessions, each with 900 trials. This program aimed to enhance the integration of visual information from the eyes and mouth regions for categorizing computer-generated faces. While C.C. demonstrated some progress on the specific tasks practiced during training, this improvement did not generalize to the recognition of novel, untrained faces.

This limited transferability highlights a potential difference in rehabilitation outcomes between acquired prosopagnosia (AP) and developmental prosopagnosia (DP). DeGutis et al. [116] reported that a less intensive version of the same training program yielded benefits for individuals with DP, improving both face perception and subjective recognition abilities. These contrasting results necessitate further research into rehabil-

itation strategies for AP to better understand the potential challenges associated with treating acquired forms of prosopagnosia compared to developmental casesd.

Moereover,study [117] examined the effectiveness of a perceptual learning program in improving face discrimination abilities in individuals with developmental prosopagnosia. Ten subjects underwent several months of training, which involved discriminating shapes between morphed facial images focusing in the core features , with the difficulty adjusted to each subject's perceptual threshold. The training progressed from neutral faces in frontal view to increasing variations in view and expression. Five subjects completed an 11-week control television task before undergoing training, while the other five were reassessed three months post-training to evaluate the maintenance of benefits.The results showed that perceptual sensitivity for faces significantly improved after the training, whereas no improvement was observed following the control task. Notably, the improvement generalized to untrained expressions and views of the faces, and there was some evidence of transfer to new faces. The benefits of the training were maintained over a three-month period.

The study involved EM [97], an adolescent with acquired prosopagnosia following encephalitis. Initial assessments revealed significant difficulties in face perception and recognition, with EM avoiding inner facial features like eyes, nose, and mouth, instead focusing on outer features such as hair and jawline. This avoidance was consistent across different emotional expressions, suggesting a generalized perceptual deficit rather than a targeted avoidance of specific facial features.

Post-training evaluations indicated substantial improvements in EM's face perception skills following a 14-week online perceptual training program. Eye-tracking data showed that EM spent significantly more time examining inner facial features post-training, aligning her viewing patterns more closely with those of control participants. This improvement extended to untrained faces, suggesting that the training enhanced EM's general face-specific processing mechanisms. Despite these gains in perceptual skills, EM's ability to recognize newly encoded faces did not improve significantly, though her confidence in social interactions increased [97].

In the Davies et al. [118] study, the focus was specifically on participants with acquired prosopagnosia. Out of the 11 participants, 8 showed improvements in their face recognition abilities following the perceptual training program, while 3 participants did not exhibit significant improvements. The effectiveness of the training varied among individuals, likely influenced by factors such as the type and severity of prosopagnosia, as well as age.

Table 2.3 provides a detailed overview of various studies conducted to assess the impact of different training durations and age on improving face recognition skills in prosopagnosia patients. Of the 24 cases presented, 17 reported significant improvements in face recognition abilities, while 7 cases did not show notable enhancements. The studies included a diverse range of participants, both in terms of age and gender, with the duration of training programs varying widely from 14 days to 18 months.

Short-term training programs (under one month) generally showed mixed results. For example, [111] reported improvements after just 14 days, while [109] and C.C. in [115] showed no significant improvement after 3 weeks and one month respectively. Medium-term training programs (1-6 months) appeared more effective, with [112] and [107] both reporting improvements within this timeframe. Long-term training programs (over 6 months) consistently showed positive results, with [113] and [108] reporting significant improvements after 14 and 11 months respectively.

The age of the patients also played a role in the outcomes. Younger patients (under 20 years) showed varied responses. While K.D. (8 years) in [110] and T.M. (12 years) in [114] did not improve significantly, A.L. (8 years) in [107] and B-ATOT3 (15 years) in [118] showed marked improvements. Middle-aged and older patients (20-60 years) generally responded well to training, with several cases showing improvement after sustained training efforts. For instance, N.E. (21 years) in [111] and M.Z. (48 years) in [113] both demonstrated significant improvements. Very young children (around 4 years), such as K. in [106], showed promising improvements, suggesting that early intervention can be beneficial.

In summary, as shown in Table 2.3, 17 out of 24 cases reported significant improvements in face recognition abilities through targeted training programs, particularly among individuals with acquired prosopagnosia. Conversely, 7 cases did not show notable enhancements, highlighting the variability in response to rehabilitation efforts. Factors such as the patient's age, type of prosopagnosia, and specific brain regions affected appear to influence the effectiveness of these interventions. These findings suggest that while many individuals can benefit from structured training programs, further research is necessary to optimize these approaches and address the needs of those who do not respond to current rehabilitation methods. Looking ahead, our research will investigate the potential of applying this focused approach to enhance deepfake technology detection. Specifically, we aim to determine whether emphasizing central facial features in deepfake models could lead to breakthroughs in identifying deepfakes more effectively.

2.6 Computational Neuroscience and Prosopagnosia

2.6.1 Facial Recognition in the Human Brain

In the human brain, face recognition primarily involves the fusiform face area (FFA), which is located in the inferior temporal cortex. Notably, damage or dysfunction in this region can result in prosopagnosia, a condition characterized by the inability to recognize faces [119].

Furthermore, the brain utilizes a combination of hierarchical processing and experience-driven plasticity to learn and recognize faces. Specifically, neurons in the FFA exhibit selective responses to faces, and through repeated exposure, these neural circuits become more efficient at distinguishing individual faces [120].

Additionally, synaptic plasticity, the capacity of synapses to strengthen or weaken over time, is crucial for learning. Hebbian learning principles, often summarized as 'cells that fire together wire together,' are central to this process [121]. This principle explains how simultaneous activation of neurons leads to the formation of stronger synaptic connections, thereby enhancing the brain's ability to recognize and differentiate faces [122].

2.6.2 Facial Recognition in Prosopagnosia

The brain areas crucial for normal face recognition, particularly the fusiform gyrus located in the temporal lobe, may be damaged or underdeveloped in people with prosopagnosia. This area is responsible for processing and storing facial information [119].

Moreover, individuals with prosopagnosia often struggle to encode and remember the specific features and configuration of faces, making it difficult to build a mental representation for identification [120]. Consequently, they may rely more on non-facial cues like hairstyle, clothing, voice, or context to recognize individuals [123].

Fortunately, the brain exhibits a degree of plasticity, allowing it to reorganize itself to some extent. Therefore, training programs for prosopagnosia leverage this plasticity by providing repetitive exercises that focus on specific facial features and their arrangement. These programs aim to help individuals with prosopagnosia develop new strategies for face recognition, even if they do not fully restore normal function [124].

2.6.3 Facial Recognition in Artificial Intelligence

AI systems, particularly convolutional neural networks (CNNs), are designed to emulate the hierarchical structure of the visual cortex. Specifically, these systems use layers of artificial neurons to process and recognize patterns in data, much like the visual cortex processes visual information [125].

To begin with, these AI systems learn through three primary methods: supervised, unsupervised, and reinforcement learning. In supervised learning, the network is trained on labelled data. Consequently, the system adjusts its weights through backpropagation to minimize error. This process involves comparing the network's predictions to the actual labels and iteratively adjusting the weights to reduce discrepancies [126].

Furthermore, although not biologically based, AI systems exhibit a form of plasticity through the adjustment of weights. Gradient descent algorithms iteratively adjust these weights to improve performance, allowing the system to adapt and refine its pattern recognition capabilities [127].

Moreover, AI systems require vast amounts of data to achieve high accuracy in face recognition. This extensive data is used to train models, enabling them to recognize faces with remarkable precision [128].

In addition, layers in CNNs progressively extract features from raw pixel data, mirroring how the visual cortex processes visual information. Early layers in the network detect basic features such as edges and textures. As the data moves through deeper layers, the network identifies more complex patterns, including facial features [129].

Finally, well-trained AI models can achieve performance levels that are comparable to or even surpass human capabilities in face recognition tasks. This high level of accuracy results from extensive training on large datasets and the sophisticated architecture of CNNs, which allows for detailed and nuanced pattern recognition [130].

2.6.4 Similarities and Differences in Face Recognition Between the Human Brain and Artificial Intelligence

Similarities

Hierarchical Processing : Both biological and artificial systems use hierarchical processing to decompose complex visual information into simpler components for analysis [125, 131].
Experience-Driven Learning : Just as the brain relies on repeated exposure and experience to fine-tune face recognition, AI systems depend on large datasets and iterative training to improve accuracy [129, 120].
Adaptation and Plasticity : Both systems exhibit forms of plasticity, with biological systems adjusting synaptic strengths and AI systems adjusting neural network weights [121, 127].

Differences

Biological Complexity vs. Artificial Simplification : The human brain's mechanisms are more complex and nuanced, involving a multitude of interconnected regions and processes beyond simple neural activation. AI models, while inspired by biological systems, simplify these processes to make computation feasible [126].
Data Requirements : AI systems typically require much larger amounts of data for training compared to the human brain, which can learn from fewer examples through more efficient generalization mechanisms [132].
Learning Processes : The human brain uses a combination of supervised and unsupervised learning inherently, with reinforcement from real-world feedback. In contrast, AI models often rely on explicit supervised learning with predefined datasets [133].

2.6.5 Parallels Between Prosopagnosia and Deepfake Detection

The thesis explores the parallels between the cognitive processes in prosopagnosia and AI learning mechanisms:

- Feature Importance: Human cognitive systems and artificial intelligence (AI) models both heavily rely on certain facial features for recognition. Crucial

for identifying faces, the eyes and nose are central to both human perception and AI feature extraction processes. Recent advancements have shown that Convolutional Neural Networks (CNNs), specifically designed for face recognition, have reached human-level accuracy [134, 130]. Studies further support that both humans and AI systems primarily use the same internal features, including the eyes, nose, and mouth, to recognize faces effectively [134, 130].

- Training and Improvement:Human cognitive systems and artificial intelligence (AI) models significantly depend on distinct facial features for face recognition. In humans, essential features such as the eyes and nose are crucial for recognizing faces, a focus paralleled in the feature extraction processes utilized by AI models. Convolutional Neural Networks (CNNs), specifically optimized for face recognition, have reached human-level accuracy, as evidenced by recent developments [135]. Subsequent studies affirm that both humans and AI systems primarily rely on the same internal facial features—namely, the eyes, nose, and mouth—for effective face recognition [134]. Moreover, this reliance on specific facial features underpins training programs aimed at improving the face recognition capabilities of individuals with prosopagnosia by focusing on the eye and nose regions [98, 112, 106, 107].Therefore, further research is warranted to explore how effectively facial cues employed in prosopagnosia training programs could enhance the performance of deepfake detection models, potentially contributing to advancements in both human and machine learning domains in face recognition.

Techniques inspired by research on prosopagnosia, such as varying the focus on different facial features, can enhance the robustness of AI models by introducing variability in the training data. This method reduces overfitting and improves the models' capacity to generalize from the training set to new, unseen images. A common technique involves partially obscuring images in the training dataset to prevent overfitting—a problem prevalent in datasets containing deepfakes [136].

The AI learning mechanism outlined in this thesis utilizes deep neural networks to refine deepfake detection strategies by incorporating principles from training and rehabilitation programs aimed at enhancing facial recognition abilities in individuals with prosopagnosia. These programs focus training on critical facial features, sometimes leading to significant improvements. By applying similar techniques, this research aims to train AI models to distinguish between real and fake images effectively, thus deepening our understanding of the role facial features play in deepfake detection.

This thesis employs advanced deep neural networks, specifically EfficientNet-B7 and XceptionNet, which are trained on datasets such as Celeb-DF and FF++. Face cut-out augmentation techniques are also used to obscure different facial regions to assess the impact of such obscuration on model performance.

In summary, this thesis integrates deep learning models with innovative augmentation techniques derived from medical research on prosopagnosia to increase the accuracy and robustness of deepfake detection. This interdisciplinary approach combines technological advancements with medical insights to tackle the challenges posed by deepfake technologies. Further experimental details will be provided in Chapter 3.

2.7 Analysis and Interpretation of Literature Survey Findings: Prosopagnosia Research and Deepfake Technology

The literature survey in this chapter explored the intersection between deepfake technology and the medical condition known as prosopagnosia to investigate the potential of applying insights and strategies from the medical field to improve deepfake detection models.

The rationale for linking these two domains lies in their similarities, particularly in the principles of hierarchical processing, learning mechanisms, and plasticity, as detailed in Section 2.6. Our review of deepfake detection methods revealed that techniques employing biometric cues, especially those focusing on the eye regions of the face, achieved higher accuracy compared to other methods. These cues prove more resilient because they are difficult to mimic due to their unique details.

Menotti et al. [62] adopted a comprehensive approach by employing the iris as a cue to detect deepfakes. This method resulted in a remarkable accuracy of 98.93%, the highest among all methods evaluated in Section 2.4.4. Another study by Jung and Jun [59] introduced a technique for identifying deepfakes by focusing on eye blinking as a key indicator, achieving an accuracy rate of 87.5%.

Similarly, a study by Xin Yang et al. [61] focused on central facial features, particularly the eyes and nose, achieving an accuracy of 95.5% in detecting deepfake images. This research utilized the differences between head poses estimated using central facial landmarks and those in the central face regions. Specifically, it exploited

directional discrepancies in the nose region to distinguish deepfakes from genuine images effectively.

In contrast, a method by Lee et al. [137] for deepfake detection, which utilized the mouth and teeth as cues, reported a lower accuracy of 71.25%. Notably, approaches centered around the eye region tend to yield higher accuracy compared to those relying on the mouth. To enhance precision in detecting deepfakes, one viable strategy could be the integration of multiple cues to improve deepfake recognition accuracy.

These studies focus on using central facial features to indicate deepfakes, highlighting that different facial features play varying roles in distinguishing between real and fake faces.

Building on these findings, face recognition research also emphasizes the crucial role of the eyes in various face-related tasks.tier, Villate, & Ryan [138]; Henderson, Williams, & Falk [82] found that these tasks include significantly influencing identity and emotion recognition, as well as the ability to understand others' mental states. However, Conditions such as prosopagnosia and Autistic Spectrum Disorders (ASD) involve difficulties in processing facial configurations [139]. Individuals with these conditions typically avoid eye contact and struggle to read mental states from the eyes alone [140].

In Section 2.5, we delve deeper into prosopagnosia, examining the challenges individuals face in recognizing differences between faces and how training and rehabilitation programs are designed to improve their ability to distinguish faces. As outlined in Section 2.5.6, case studies summarize the detection cues relied upon by individuals with prosopagnosia. Notably, these cues often exclude the eye region, focusing more on external facial and non-facial cues such as hairstyle and skin tone [97, 103, 102, 99].

Section 2.5.7 reviews strategies employed by medical experts to enhance the facial identification abilities of individuals with prosopagnosia. While most cases show improvement following training, some do not. However, Mann et al. [141] state that "training can be used to enhance facial recognition capabilities." In Table 2.3, nine studies indicate that training enhances the ability to identify faces, whereas four studies report no improvement. These variations may be attributed to differences in each case of prosopagnosia, such as the area of brain damage, cognitive abilities affected, age, and gender factors.

All these training programs focus on shifting the fixation and attention of individuals with prosopagnosia towards the central features of the face, especially the eyes and

nose regions—areas they typically avoid. This approach is based on the established understanding within face recognition research that central facial features play a critical role in effective face recognition.

Our deepfake detection model will be trained to focus on specific facial regions using novel face cut-out techniques, which will be detailed in Chapter 3 . These cut-out regions are strategically selected based on insights derived from prosopagnosia research, which have identified critical facial features essential for face recognition. By obscuring different parts of the face during training, the model will be compelled to concentrate on the remaining visible regions, thereby enhancing its ability to detect deepfakes. This approach leverages the understanding that prosopagnosia patients can improve their face recognition skills by focusing on key facial features, such as the eyes and nose. The application of these principles aims to improve the robustness and accuracy of deepfake detection systems by emphasizing the most informative facial regions during the training process.

In conclusion, based on insights gained from the medical field, particularly the condition of prosopagnosia, we plan to refine our deep neural model for detecting deepfake images by focusing on specific facial regions that are emphasized in prosopagnosia training programs. To implement this, we will utilize innovative face cut-out techniques, which are elaborated upon in Chapter 3. These cut-out regions are strategically selected based on research findings from prosopagnosia studies, which have pinpointed critical facial features essential for accurate face recognition.

By selectively obscuring parts of the face during the model's training phase, we aim to compel the model to focus on the remaining visible regions. This method enhances the model's ability to detect deepfakes by training it to recognize subtle discrepancies in the most informative facial areas, such as the eyes and nose. This training approach is inspired by the adaptation strategies employed by individuals with prosopagnosia, who improve their face recognition abilities by concentrating on key facial features. Ultimately, the application of these principles is intended to significantly improve the robustness and accuracy of deepfake detection systems by emphasizing the facial regions most critical for identification during the training process.

2.8 Related Works

The present study covers the scientific areas of medicine and information technology. Strategies from the medical field are used, particularly from information on prosopagnosia, to improve deep fake detection methods. Data augmentation for images, which shares some similarities with this work, is also discussed. This section is divided into three parts which consider deepfake detection methods, deepfakes and medicine, and data augmentation.

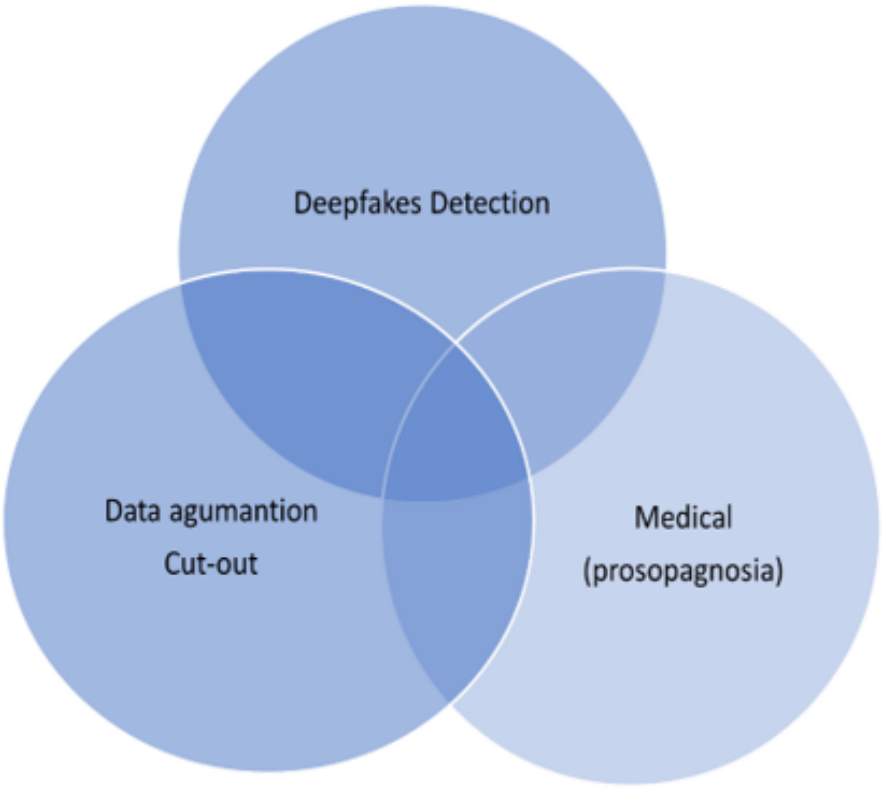


Figure 2.4 Research Overlaps

2.8.1 Deepfakes Detection Methods

This section discusses various techniques used for detecting deepfakes, which are manipulated media files generated by AI algorithms, focusing on methods that use biometric features to detect fakes. Researchers have explored the use of biometric

features such as facial landmarks, eye movement, and patterns of blinking to detect deepfakes.

Several studies have been conducted to identify the artifacts left by deepfake generators, including face warping [142], temporal and spatial inconsistencies [57], eye blinking [68], inconsistent head poses [61], and others. In one study, various experiments were performed to determine the most effective image features for the detection of fake faces in general [143]. Furthermore, a survey comparing the performance of multiple deepfake detection architectures has recently been published [144]. The features of the face, such as the eyes and mouth, are now commonly used to identify deepfakes, and these approaches achieve good levels of performance.

Research conducted by Jung et al. [59] and Li et al. [68] offers interesting approaches to the detection of deepfake videos by focusing on patterns of the blinking of the eyes. Blinking is a spontaneous and involuntary action, and there are predictable patterns of eye blinking that can be used to identify deepfake videos.

The deep vision algorithm developed by Jung et al. [59] was able to successfully detect seven out of eight deepfake videos. However, it is important to note that this approach is only effective if the deepfake video involves a person who blinks naturally and does not exhibit any unusual blinking patterns or abnormalities. While this study focused on using the single cue of blinking patterns to detect deepfake videos, it is possible that combinations of multiple cues could improve the accuracy of detection. For example, features such as lip movements or facial expressions could be analysed in combination with blinking patterns to increase the reliability of deepfake detection. Overall, the analysis of eye blinking is a promising step towards detecting deepfake videos, but more research is needed to determine its effectiveness in real-world settings and to explore the potential of using combinations of cues for deepfake detection.

Menotti et al. [62] focused on the development of a method for detecting spoofing, which is the act of impersonating someone or something, using different modalities such as the face, iris, and fingerprint. The authors propose a convolutional network architecture that performs architectural and filter optimisation for the purposes of identification. Additionally, they note that disparities in iris movement can be used as a clue for detection to distinguish between fake and real videos. Overall, positive results are reported in resolving various problems related to the detection of spoofing.

One potential extension to Menotti et al.'s work could be to explore the effectiveness of incorporating other biometric data sources, such as those concerning the voice or gait, in order to further enhance the accuracy of detection achieved by the system. Overall,

Menotti et al.'s work is a valuable contribution to the field of biometric security and has the potential to inform the development of more effective anti-spoofing systems.

The method and techniques used by Jafar et al. [53] were able to accurately detect differences in mouth movements between real and deepfake videos. This study used the Deepfake Forensics (Celeb-DF) and Deepfake Vid-TIMIT datasets, which are commonly used datasets in deepfake detection research, and focused on the analysis and comparison of the width and openness of mouth movements when certain words were pronounced. The results suggest that mouth movements in fake videos are wider and more open than in real videos, providing a potential clue for the detection of deepfakes. However, it is important to note that deepfake technology is constantly evolving, and new techniques may emerge that could potentially circumvent this type of detection.

The primary objective of the present research is to improve our understanding of facial features and to identify the regions that provide the most informative content for the purpose of detecting deepfakes.Additionally, the potential benefits of incorporating multiple regions simultaneously are also investigated.The selection of these facial regions is guided by medical insights, particularly those gained from the study of prosopagnosia. Through a systematic approach, this work seeks to improve the accuracy and robustness of deepfake detection systems.

2.8.2 Deepfakes and Medicine

This section examines the relationship between deepfake technology and the medical domain. A growing body of literature has demonstrated potential applications of deepfake technology in various aspects of medicine, and there has been a notable increase in reliance on artificial intelligence (AI) technology.

Deepfake technology, which utilizes deep learning algorithms to manipulate images, has shown promising potential in enhancing medical image analysis and diagnosis [145]. One potential application involves the creation of synthetic medical images, which can be used to train machine learning models and improve their accuracy in the detection of diseases. For example, researchers have used deepfake technology to create synthetic X-ray images, which can help improve the accuracy of automated X-ray diagnosis.

Additionally, deepfake technology can be used to simulate medical procedures, providing medical students and professionals with realistic training scenarios. This can help improve their skills and enhance patient safety. For instance, virtual reality

simulators can be created using deepfake technology to simulate complex medical procedures such as surgeries, enabling medical professionals to gain experience and proficiency before performing them on real patients [146].

Falahkheirkhah et al. [91] presented a novel approach to generating histologic images using a generative adversarial network model. This approach not only reproduces the diagnostic morphological features of common diseases but also allows users to generate new and rare morphologies. The framework was tested on synthetic data for prostate and colon tissue images and was found to be useful in augmenting the diagnostic ability of machine learning methods. The usability of the images by a panel of experienced anatomic pathologists was also assessed, and it was found that the pathologists were not able to distinguish between real and synthetic images.

Moreover, the analysis showed a similar level of interobserver agreement for prostate cancer grading. The approach was also extended to significantly more complex images from colon biopsies, and the morphology of the complex microenvironments in such tissues was reproduced. Finally, the study demonstrated the ability for a user to generate deepfake histological images using a simple markup of semantic labels.

Deepfake technology has been shown to be effective in addressing data scarcity in certain applications, such as medical imaging. By generating high-quality synthetic images that mimic real images, deepfake technology can help expand the size and diversity of available datasets, which can, in turn, improve the performance of deep learning models. In the case of knee imaging [147], for example, if there are only a limited number of real knee images available for the training of a deep learning model, it may be difficult for the model to learn to accurately segment different structures within the knee. However, by using deepfake technology to generate additional synthetic knee images, the model can learn from a larger and more diverse dataset, which can improve its ability to accurately segment knee structures.

The prominent role of big data in modern medical science has been underscored by recent global developments. However, the issue of privacy constitutes a major hurdle in the collection and sharing of data between researchers. To overcome this challenge, a recent study [148] presented synthetic data using deepfake technology, which was generated to mimic real data while carrying similar information and distributions. By using synthetic data, privacy breaches can be avoided, and the confidentiality of patients' sensitive information maintained, while researchers are still able to continue their work. This approach can not only benefit the medical community but also has broader implications for other fields that rely heavily on big data analysis.

Another study [149] has addressed the privacy issues surrounding videos of patients. The sharing of medical research data has always been challenging due to privacy concerns about clinical data since its open-sourcing may potentially violate the privacy of the patients involved. Traditional methods of face de-identification, such as blurring or pixelation, however, remove all facial information and render it impossible to analyse facial behaviour. However, recent advances in the detection of whole-body key points have demonstrated the critical role that facial information plays in estimating these key points accurately.

In some medical diagnoses, both facial and body key points are crucial, and maintaining the invariability of these key points after de-identification is of great importance. As a potential solution, one study [149] has proposed the use of deepfake technology and the face-swapping technique to protect patient privacy in medical videos. While the use of this technique has been criticised for invading privacy and violating portraiture rights, it could nevertheless be used to protect privacy in medical videos. Using this technique, the faces of patients can be swapped for suitable target faces, rendering them unrecognizable while preserving important facial and body information for medical diagnosis and research purposes. By leveraging deepfake technology, researchers can maintain the accuracy and reliability of their data while safeguarding patient privacy.

Another study [150] has presented a practical and lightweight technique used to accelerate deepfake detection in biomedical imagery by detecting malignant tumours in the modalities of healthy patients. The technique is based on convolutional reservoir networks (CoRNs), which enable ensemble feature extraction and result in improved classification metrics. This approach has the potential to significantly enhance the accuracy and reliability of medical imaging analysis and could ultimately benefit medical diagnosis and research.

The use of deepfake technology in the medical field has increased in recent years, providing a variety of benefits, such as increasing the size of medical datasets, protecting patient privacy, and improving diagnoses. However, there has been limited research on how strategies originating in the medical field could be used to enhance or improve deepfake detection technology. The present research aims to bridge this gap by leveraging medical expertise concerning the condition of prosopagnosia. Specifically, this research will leverage medical strategies that are used to help patients with prosopagnosia identify differences between faces. By using these strategies, the aim is to develop a more accurate and reliable method for the detection of deepfakes. The

understanding and insights derived from medical research on prosopagnosia are thought to have substantial potential in enhancing deepfake detection techniques.

2.8.3 Deepfakes and Data Augmentation

Data augmentation is an essential technique used to improve the performance of convolutional neural networks (CNNs) in image recognition tasks. This process entails implementing a range of alterations to the input images during training, including actions like flipping, rotating, randomly cropping, jittering, translating, injecting noise, altering colours, and more, as detailed in Shorten et al.'s survey [136].

One popular data augmentation technique is random erasing, where random patches from the input image are cut out or replaced with noise during training. This method helps the network to learn or estimate features by matching neighbouring information in the image, which can improve the robustness of the model. However, randomly cutting out patches can result in the removal of essential object descriptors, which could be detrimental to the training process [151].

Therefore, selecting the appropriate data augmentation techniques is crucial to prevent eliminating vital object descriptors or introducing bias into the training dataset. Moreover, vigilant monitoring of the model's performance throughout the training phase is essential to identify any emerging issues and to modify the augmentation strategies as needed [136].

The cut-out is a simple regularisation technique used for convolutional neural networks that involves removing contiguous sections of input images. This technique serves multiple purposes, including increasing the effective size of the training dataset by creating modified copies of original images, as well as biasing the model so that more attention is paid to specific regions of the images [152]. The position and size of cut-out regions can be randomly determined during training and may be applied to any region of the image, or limited to specific regions such as the face in a facial recognition application.

Research has demonstrated that the straightforward regularisation technique of randomly masking out square regions of input data during the training phase, referred to as cut-out, can enhance the robustness and overall performance of convolutional neural networks [152]. This technique is remarkably simple to implement and can be used alongside existing forms of data augmentation and other regularisers to further enhance a model's performance.

Random cut-out has been used to improve image classification, object detection, and person re-identification in deep learning models [151]. There has recently been an increase in the use of cut-out techniques, especially in digital images featuring faces. A recent study [153] explored the use of face cut-out and random cut-out augmentations. These were separately applied to train two different models for use in the detection of deepfakes. These cut-out augmentations were used to prevent overfitting, which is a common problem in machine learning where a model becomes too specialised to the specific training data and fails to generalise well to new data. One limitation of this study is that the model's ability to generalise to new data could be improved by using previously untapped data and applying it to multiple datasets. However, another study [154] tried to solve the generalisation problem in most deepfake detection models, so that their proposed model would be sensitive to different types of forgeries by using a large forgery augmentation space. This study further proposes the use of the adversarial training strategy and adversarial data augmentation to dynamically synthesise the types of forgeries most challenging to the current model.

In addition to preventing overfitting and improving generalisation, another function of the face cut-out technique in deepfake detection has been to create a dataset consisting entirely of masked faces [155]. The motivation for this study was the need during the Covid-19 pandemic to train deepfake detection models to recognise manipulated faces even when certain areas of the face, such as the mouth and nose, were covered by a mask. To address this, the authors generated both real and fake faces wearing masks to create a test dataset for the evaluation of approaches to deepfake detection. By training the model with both real and fake masked faces, they were able to improve its ability to detect deepfakes with masked regions. One approach used to address the overfitting problem in deepfake detection involved the use of two cut-out operations: sensory group removal and Convex-hull removal [156]. Our experimental technique was inspired by Sowmen Das et al.'s solution [156], which addresses the overfitting problem in deepfake detection using two cut-out operations: sensory group removal and Convex-hull removal.

- Sensory Group Removal: This technique randomly selects one of three landmark groups (two eyes, nose, and mouth) and removes the maximum polygonal region defined by the group's points.
- Convex-hull Removal: This method involves selecting landmark points to represent the facial boundary and calculating the maximum polygon generated by the points with the minimum envelope. Points can be randomly selected from

all boundary points or as a number of contiguous points having the maximum polygonal area with the minimum envelope. Alternatively, points can be selected from one of four sub-polygons created by partitioning the outer polygon.

This study compares selective and extensive cut-out techniques to determine which approach improves model performance in detecting deepfakes. However, the potential drawback of the face-cutout method is the lack of a clear explanation regarding the selection of cut-out regions, particularly why certain facial parts are removed. In this study, the first group applied the face cut in a Convex-hull shape, covering half of the face and multiple regions at once. The second group used a more selective face cut on three facial areas: eyes, nose, and mouth. This specificity helps to understand the impact of these regions on model performance. However, this method does not consider the role of external facial features, which may contain crucial information for detecting deepfakes. Most deepfake creation processes focus on altering internal facial features while leaving external features unchanged, potentially aiding in deepfake detection.

To address this issue, the present research focuses on discovering which facial features are most helpful for deepfake detection by applying the face cut in two separate groups:

- Internal Regions: Covering both eyes, right eye, left eye, and nose.
- External Facial Features: Covering chin, mouth, jawline, and forehead.

The selection of these facial features in each group is inspired by medical research, particularly training programs designed to improve face recognition in individuals with face blindness (prosopagnosia).

In related work, it has been observed that the utilisation of AI in the medical field has recently increased significantly. AI has facilitated processes traditionally time-consuming when performed by humans, such as diagnosing certain diseases [145], simulating medical procedures, and providing realistic training scenarios for medical students and professionals [146]. Furthermore, AI has proven to be beneficial in preserving privacy and addressing data scarcity in specific applications like medical imaging by generating high-quality synthetic images that closely mimic real ones [91].

Conversely, there is a notable paucity of studies leveraging medical knowledge to enhance AI, particularly in the realm of deepfake detection. Research in deepfake detection and creation frequently recycles existing methodologies, leading to a perpetual

competition between those detecting deepfakes and those creating them. To disrupt this cycle, we propose utilising insights from the medical field, specifically from the study of prosopagnosia.

Our objective is to identify which facial features provide the most informative content that aids prosopagnosia patients in improving their face recognition abilities. We aim to determine if these features can similarly enhance AI models in recognising deepfakes.

To achieve this, we will modify the dataset used to train the model, focusing on these specific facial features. We will employ data augmentation techniques, particularly cut-out techniques, initially used to mitigate overfitting by randomly placing cutouts within training dataset images [151]. Recently, studies have started using selective face cutouts to cover specific facial areas. For instance, some studies have masked the nose and mouth to create datasets that train models to recognise deepfakes even when faces are masked [155]. Other studies have compared selective cutouts with extensive cutouts, where half of the face is covered, to determine which method improves model performance [156].

Thus, similar techniques have been applied to different challenges and facial areas. In our research, we will utilise selective face cut-out techniques to cover specific facial features, compelling the model to learn from these areas and evaluate their impact on detecting deepfakes. The selection of these cutouts is informed by insights gained from rehabilitation and training programs that assist prosopagnosia patients in improving their face recognition abilities. For our experiment, we have devised two groups of cutouts: one focusing on internal features and the other on external features.

Overall, integrating insights from various disciplines and developing clearer, more interpretable methods for dynamic face augmentation is crucial for advancing deepfake detection in images. This approach aims to enhance the reliability and accuracy of deepfake detection in images.

2.9 Chapter Summary

This chapter explores the intersection of deepfake detection and prosopagnosia, drawing on a survey of the literature to identify overlaps between these two fields. It provides valuable insights into the research conducted in both disciplines, with the primary goal of exploring connections between deepfake recognition and face recognition disorders.

Given the extensive studies on prosopagnosia, the chapter applies relevant findings from this field to the challenges posed by deepfakes. The identification of critical facial features plays a key role in improving facial recognition techniques, which can help counter the issues associated with deepfakes. Additionally, this exploration may offer valuable directions for future research by bridging these distinct disciplines.

Moreover, the chapter integrates scientific knowledge from medicine and information technology, using medical strategies—specifically insights from prosopagnosia—to refine deepfake detection methods. It also discusses data augmentation techniques for images and suggests how medical understanding of facial recognition disorders can enhance the training of deep learning models for more accurate deepfake identification. This interdisciplinary approach not only deepens our understanding of deepfakes but also paves the way for future research, potentially leading to more effective solutions against digital misinformation.

Chapter 3

Experiment Plan and Setting

3.1 Introduction

The primary aim of this study is to explore the intersection between prosopagnosia, a medical condition impairing face recognition, and the field of deepfake detection. This research seeks to determine if medical insights into face recognition can enhance deepfake detection techniques. To achieve these objectives, the dissertation employs a multifaceted methodology, focusing on identifying key facial features crucial in differentiating authentic faces from deepfakes and understanding how prosopagnosia affects this recognition process.

The methodology of this study is organized into seven stages, each designed to contribute to a comprehensive understanding of both prosopagnosia and deepfake technology. These stages range from literature review and technology assessment to practical experiments and data analysis. The approach aims to bridge the gap between medical research and technological challenges in digital media, potentially leading to innovative strategies in deepfake detection.

3.2 Identification of Cues

This stage involved an in-depth review of existing literature on prosopagnosia, with the aim of applying insights from this research to the domain of deepfake analysis, as elaborated in Chapter 2. The focus was on identifying key facial features, which are crucial for the accurate reproduction of facial images and could potentially enhance

facial recognition methods for countering deepfakes. The review entailed a thorough exploration of the unique characteristics associated with prosopagnosia and facial recognition, including aspects of face processing and eye movement patterns specific to individuals with prosopagnosia.

Additionally, the study examined the cues critical for facial identification and discussed the effectiveness of training sessions in improving face processing abilities in prosopagnosia patients. This comprehensive analysis aimed to develop a nuanced understanding of facial recognition mechanisms, both in individuals with typical development and those affected by prosopagnosia.

The exploration of available studies on prosopagnosia was a crucial step in our research, as it provided valuable insights into the mechanisms underlying facial recognition. By identifying critical facial features and understanding how they are processed in the brain, we gained a better understanding of how AI can be used to reproduce faces with greater accuracy.

A particularly intriguing finding from the review highlighted that individuals with prosopagnosia predominantly rely on external facial features and physical characteristics to recognize other. Thus, most current rehabilitation and training programmes focus on changing patterns of eye movement and encouraging individuals with prosopagnosia to concentrate on internal facial features. This can be effective in improving their ability to recognize faces and navigating social interactions [112] [106] [113].

The present research focuses on improving the accuracy of deepfake detection by utilizing the same techniques used in medical research to improve face recognition in individuals with prosopagnosia. By selectively blocking external or internal facial features and training the model employed to focus on specific facial features, the aim is to improve the model's ability to accurately detect deepfakes. Therefore, the same clues were used that prosopagnosia patients use to identify differences between similar faces, along with the strategies followed by medical staff in rehabilitation programmes to improve the ability of prosopagnosia patients to recognize faces.

3.2.1 Selection of Face Cut-out Regions

The identification of critical facial features is crucial in the accurate reproduction of faces and to reinforce facial recognition techniques in order to overcome deepfakes.

Research on prosopagnosia has demonstrated that the internal parts of the face, particularly the eyes and the nose, provide more information for recognition than external features such as the chin, mouth, and hair [108] [111] [107] [113]. Consequently, medical professionals have developed training programs to help prosopagnosia patients focus on these internal facial features to enhance their ability to identify individuals [107] [106].

There is substantial evidence suggesting that individuals with prosopagnosia can experience significant improvements through targeted training programs. For instance, Beyn and Knyazeva reported that their patient, C.H., showed some improvements in recognizing faces in real-world situations following a systematic training program focused on facial features and expressions [108]. Similarly, Schmalzl et al. found significant improvements in the attention to and recognition of internal facial features in a child with congenital prosopagnosia [106]. Moreover, Mayer and Rossion reported enhanced recognition abilities in a patient after training them to analyze internal facial features [112].

However, these techniques do not completely cure prosopagnosia but rather help to reduce its impact. Since prosopagnosia is often caused by neurological factors, it cannot be entirely cured through training. Nonetheless, these techniques can substantially enhance facial recognition capabilities. For example, DeGutis et al. reported improvements in standardized tests and practical daily life recognition tasks post-training, although these improvements did not entirely eliminate recognition difficulties [113].

It is also important to note that these training programs do not provide a universal cure and do not work equally well for everyone. The effectiveness of the training varies depending on the severity of the condition, the specific neurological impairments involved, and the individual's ability to learn and apply the techniques. For example, while some individuals with prosopagnosia, like M.Z., showed notable improvements post-training [113], others, like T.M., did not exhibit any significant improvements despite extensive training [114]. Additionally, the study by Ellis and Young found no significant improvement in face recognition for the child subject K.D. after an 18-month training program [110].

In conclusion, while these training programs can yield significant improvements in some individuals, they do not provide a complete cure for prosopagnosia and are not universally effective for all patients. However, these medical insights into prosopagnosia and the focus on internal facial features could potentially inform and improve methods

for detecting deepfakes. By understanding which facial features are most crucial for recognition, detection methods can be better tailored to identify inconsistencies and artifacts in deepfake images and videos.

In order to investigate which facial features are most informative in deepfake detection, two groups of cut-outs were created based on the findings from prosopagnosia research. Group 1 covered four external regions: the forehead, mouth, chin and jawline. The aim of this group was to train the model to focus on the internal features. In contrast, Group 2 covered the internal features, including the left eye, right eye, both eyes, and nose, in order to train the model to focus on the external features.

After training the model with the two groups separately, it was tested with an unseen test set to compare the accuracy of each group based on the discrimination of fake faces from real ones. The results of this study will help to determine which facial features are most important for deepfake detection. More details regarding the face cut-out regions are provided below.

Covering certain areas of the face in images can enhance the accuracy of deepfake detection systems. This approach aids in pinpointing the most informative facial features to concentrate on. Additionally, it assists in overcoming the overfitting problem in deepfake datasets by creating various modified versions of the original images, ensuring important regions remain uncovered. Such a method also provides valuable insights for researchers to develop more effective techniques for deepfake detection in future studies.

3.3 Face Cut-out

A face cut-out technique was employed to cover specific regions of the face as allocated in section 3.2.1. This face cut-out serves as a data augmentation method used in the training of convolutional neural networks (CNNs) to enhance deepfake detection. It generates training images with various occlusions using facial landmark information, irrespective of orientation.

Numerous libraries are available for integrating facial landmark detection into applications, including dlib, OpenCV, and PyTorch3D, among others. The decision to adopt the MediaPipe Face Mesh has been guided by its distinct advantages, notably its capability to detect and track an extensive array of 468 landmark points on the human face, providing fine-grained facial feature tracking. Furthermore, its flexibility of use,

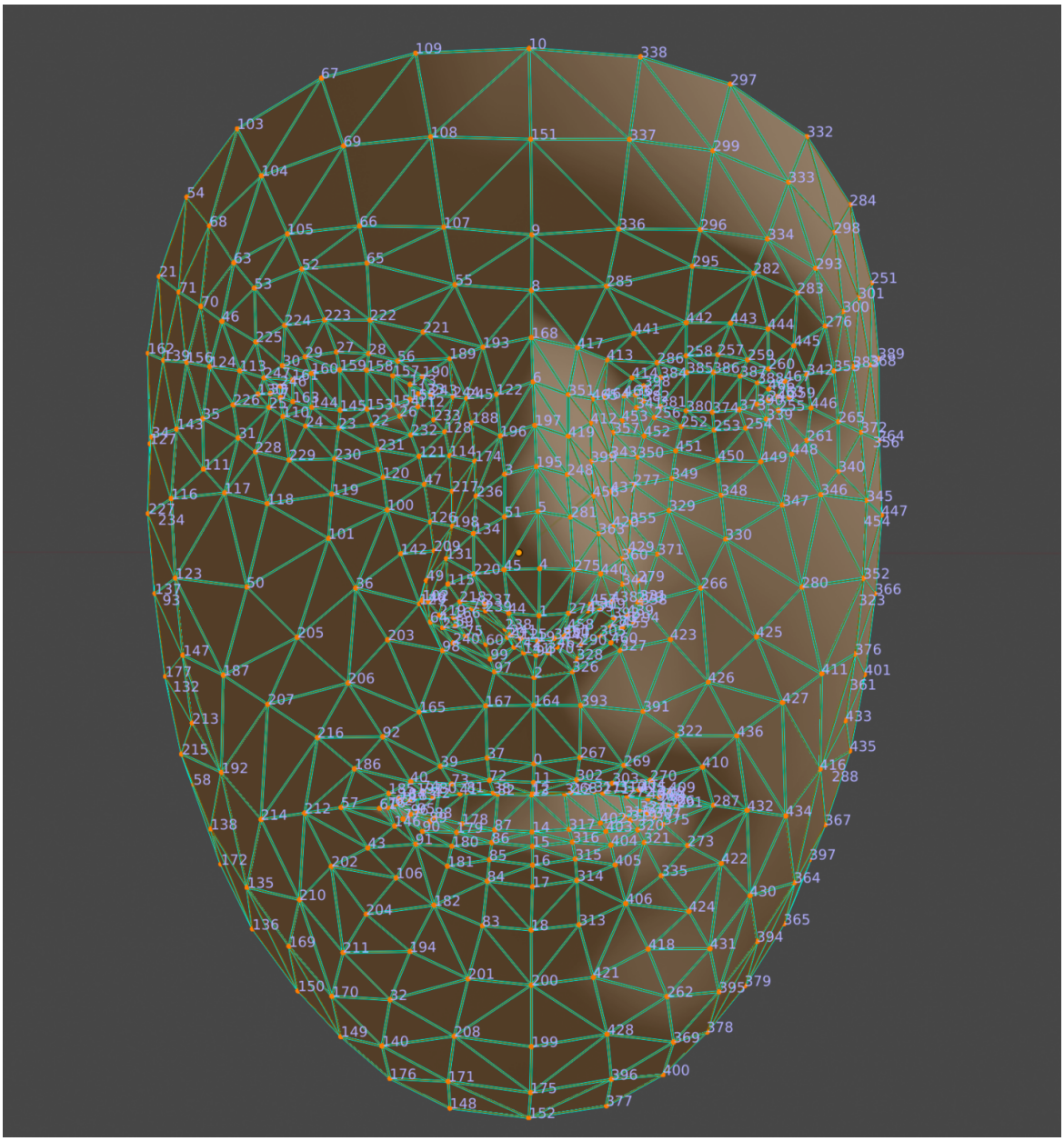


Figure 3.1 MediaPipe Face Mesh: A 3D Facial Landmark Detector with 468 Landmarks

cross-platform compatibility, and the support of an active developer community, backed by Google, ensure its ongoing refinement and reliability. These factors collectively position the MediaPipe Face Mesh as an optimal solution for our specific needs in facial landmark detection and analysis [157].

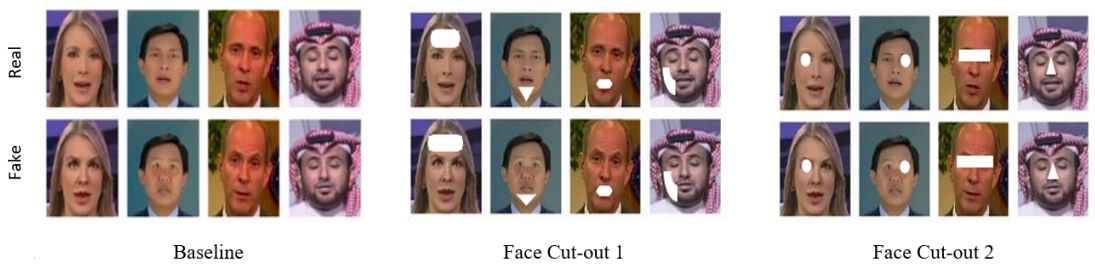


Figure 3.2 Dataset Examples from the Study: (1) Baseline Images of Original, Unaltered Faces; (2) Face Cut-Out 1 with Specific Regions Removed (Left Eye, Right Eye, Both Eyes, Nose); (3) Face Cut-Out 2 Featuring Removal of Forehead, Chin, Mouth, and Jawline.

Landmark positions on the human face encompass critical anatomical points, including the eyes, nose, mouth, jawline, and forehead. The MediaPipe Face Mesh, illustrated in Figure 3.1, is capable of discerning and uniquely identifying 468 such positions on the facial structure. Each of these landmark positions is uniquely designated with an integer ranging from 0 to 468 [157].

Importantly, the MediaPipe Face Mesh is seamlessly integrated within the MediaPipe library—a versatile, open-source, cross-platform framework [157]. This framework is specifically designed to facilitate the construction of data processing pipelines for various types of perceptual data, including video and image inputs. Within this framework, developers have access to a repository of pre-trained models, which includes not only face detection and tracking models but also models for face landmark detection and hand tracking. The availability of these pre-trained models simplifies the task of developers seeking to incorporate these sophisticated facial analysis capabilities into their applications.

To perform face cut-out, certain landmark positions were grouped together to create polygons that were then occluded in the training images. Two groups of polygons were used. The first group, Cut-out 1, consists of landmark positions for the chin, forehead, mouth, and jawline. These positions range from points numbered 211 to 150 for the chin, 103 to 67 for the forehead, 57 to 43 for the mouth, and 425 to 280 for the jawline. The second group, Cut-out 2, consists of landmark positions for the left eye, the right

eye, both eyes, and the nose. Face mesh landmark points for the left eye range from 158 to 226, for the right eye from 386 to 446, for both eyes from 53 to 340, and for the nose from 6 to 419. By using these positions to calculate polygons for the face cut-out, training images were generated with different occlusions for improved deep fake detection. Cut-outs 1 and 2 were applied to the selected datasets, as shown in Figure 3.2.

All the images were resized to a standardized resolution of 224 x 224 pixels to ensure consistency during training. The cut-out process involved several steps to ensure accuracy and effectiveness. First, landmark detection was performed using MediaPipe Face Mesh to identify 468 landmark points on the human face. Next, specific groups of landmarks were used to create polygons representing facial regions to be occluded. These polygons were then applied to the images, generating occluded versions for data augmentation.

To verify the accuracy of the cut-out regions, both visual inspection and automated verification methods were employed. For visual inspection, the output images with the cut-out regions were saved and manually inspected to ensure the correct regions were occluded. Any discrepancies or errors in the placement of the cut-outs were identified and corrected. For automated verification, the actual positions and areas of the cut-out regions were compared with the intended positions and areas based on the landmark coordinates. Metrics such as Intersection over Union (IoU) were calculated to quantify the accuracy of the cut-outs.

3.4 Experimental Set-up

3.4.1 Dataset Selection

In recent years, several deepfake datasets have been published to facilitate research and development in the field of deepfake detection. These datasets typically contain a collection of real and fake videos, where the fake videos are generated using various deepfake techniques such as face swapping and facial re-enactment.

Some of the popular DeepFake datasets include the FaceForensics++ (FF++) [158], Celeb-DF[159], and the DeepFake Detection Challenge(DFDC) [160], and they have been widely used by researchers to develop and evaluate deepfake detection algorithms and techniques.The models in this study were trained and evaluated with

FaceForensics++(FF++) and Celeb-DF, which are presently the most popular datasets from their respective generations.

FaceForensics++ [158] is a public benchmark dataset for research in to the detection of face forgery created by researchers from the Technical University of Munich and the University of Erlangen-Nuremberg in Germany and the Federico II University of Naples in Italy.The dataset is characterized by its diverse manipulations and high-quality content, which collectively challenge both human perception and algorithmic detection capabilities.It incorporates manipulations from various methodologies, including DeepFakes, Face2Face, FaceSwap, and NeuralTextures, ensuring a robust and comprehensive dataset that spans across multiple manipulation techniques. Moreover, the dataset features over 1000 original video sequences, along with their manipulated counterparts, collectively amassing a staggering total of over 5000 videos, all extracted from realistic contexts such as news interviews.

The FaceForensics++ was created to facilitate research in the area of deepfake detection and to help develop algorithms that can accurately detect manipulated videos. It has been used in several research papers and challenges, and has become one of the standard benchmarks in its field.

The Celeb-DF dataset [159], specifically its second version (Celeb-DF-v2), emerges as a pivotal resource in the domain of deepfake research, boasting a comprehensive collection of 5,639 videos, which are uniquely categorized into real and fake. The dataset encompasses videos of 32 distinct celebrities, providing a rich and diverse repository for exploring the intricacies of deepfake generation and detection. Celeb-DF, with its 590 genuine videos and a remarkable 5,049 fake videos, offers not just a wide range of content but also significant depth. This furnishes researchers with a comprehensive and diverse dataset for exploring the intricacies of deepfake technology. Additionally, it provides a balanced and realistic framework for the development and testing of deepfake detection algorithms, presenting a substantive and authentic challenge to both researchers and technologists in the field.Figure 3.3 illustrates a selection of images from each dataset used in the experiments

For this study, face cut-outs were evaluated with the FF++ and Celeb-DF datasets separately, as well as trained models with samples from the two datasets. Typically, when training a machine learning model, the data needs to be split into three sets for training, validation, and testing. The training set is used to train the model, and then the validation set is used to evaluate the model's performance during training and to make decisions about the selection of hyperparameters. The testing set is subsequently

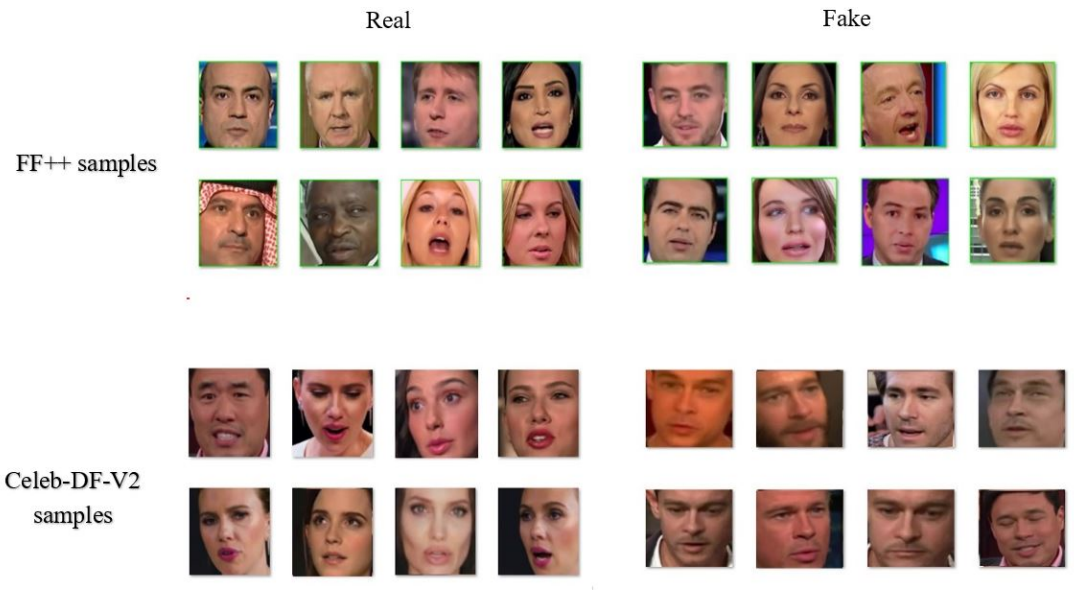


Figure 3.3 Representative Images from Each Dataset Utilized in the Study

used to evaluate the final performance of the model. In this case, 80% of the data was used for training, 10% for validation, and 10% for testing. This is a commonly employed split, but the percentages may vary depending on the size and complexity of the dataset used, among other factors.

Dataset Sizes Utilized in the Experiment

In phase one of the experiment, we trained the models using three distinct groups—Baseline, Cut-out 1, and Cut-out 2—generated from each dataset to be trained separately. Detailed information about the datasets and image counts used in Phase 1 is presented in Table 3.1.

For each dataset, the three training groups were designed as follows: The Baseline group consisted of original images without any modifications. The Cut-out 1 group included images with specific regions such as the chin, mouth, jawline, and forehead occluded. The Cut-out 2 group contained images with other regions, such as the left eye, right eye, both eyes, and nose occluded.

Table 3.1 provides a comprehensive summary of the distribution of images across these different training groups for the two datasets: FF++ and Celeb-DF v2. Each dataset is divided into three training groups: Baseline, Cut-out 1, and Cut-out 2. For

each training group, the table lists the number of fake and real images used in the training, validation, and test sets, along with the total number of images.

For the FF++ dataset, each group includes 12,800 fake and 12,800 real images in the training set. The validation set contains 1,600 fake and 1,600 real images, and the test set also contains 1,600 fake and 1,600 real images. This results in a total of 32,000 images per group. Similarly, for the Celeb-DF v2 dataset, each group includes 12,272 fake and 12,272 real images in the training set. The validation set contains 1,534 fake and 1,534 real images, and the test set also contains 1,534 fake and 1,534 real images, resulting in a total of 30,680 images per group.

This structured approach ensures a balanced and comprehensive evaluation of the models trained on these datasets. By maintaining a consistent number of images across different groups and phases, we can ensure that variations in performance are due to the specific manipulations applied to the images (e.g., cut-outs) rather than inconsistencies in the dataset sizes. This helps in accurately assessing the models' ability to detect manipulations and generalize to new data.

Dataset	Groups	Train		Validation		Test		Total
		Fake	Real	Fake	Real	Fake	Real	
FF++	Baseline	12800	12800	1600	1600	1600	1600	32000
	Cutout1	12800	12800	1600	1600	1600	1600	32000
	Cutout2	12800	12800	1600	1600	1600	1600	32000
Celeb-DF	Baseline	12272	12272	1534	1534	1534	1534	30680
	Cutout1	12272	12272	1534	1534	1534	1534	30680
	Cutout2	12272	12272	1534	1534	1534	1534	30680

Table 3.1 Summary of Dataset Size for Each Group in Phase 1

In Phase Two of the experiment, we combined the FF++ and Celeb-DF v2 datasets to enhance the model's generalization capabilities by exposing it to a more diverse set of images. This approach aimed to provide the model with a broader perspective, enabling it to better handle a variety of manipulations.

The combined dataset was divided into three training groups: Baseline, Cut-out 1, and Cut-out 2, as shown in Table 2. The Baseline group consisted of original images without any modifications. The Cut-out 1 group included images with specific regions such as the chin, mouth, jawline, and forehead occluded. The Cut-out 2 group

contained images with other regions occluded, such as the left eye, right eye, both eyes, and nose.

As shown In table 3.2 For each training group, the table lists the number of fake and real images used in the training, validation, and test sets, along with the total number of images. Specifically, each training group in the combined dataset included 25,072 fake images and 25,072 real images in the training set. The validation set contained 3,134 fake images and 3,134 real images, and the test set also contained 3,134 fake images and 3,134 real images. This resulted in a total of 62,680 images per training group.

By combining the datasets and maintaining a balanced number of fake and real images across all training, validation, and test sets, we ensured that the model was exposed to an unbiased dataset. This balanced approach was crucial for assessing the model's robustness and its ability to generalize well to a wide variety of manipulations, effectively handling diverse and previously unseen data.

Dataset	Groups	Train		Validation		Test		Total
		Fake	Real	Fake	Real	Fake	Real	
Combined Dataset	Baseline	25072	25072	3134	3134	3134	3134	62,680
	Cutout1	25072	25072	3134	3134	3134	3134	62,680
	Cutout2	25072	25072	3134	3134	3134	3134	62,680

Table 3.2 Summary of Dataset Size for Each Group in Phase 2

The table3.3 presents a comprehensive summary of the distribution of images across various training groups for the FF++ dataset during Phase 3 of the experiment. Each training group is specifically designed to focus on a particular facial region or maintain a baseline condition, with the objective of evaluating the model's performance when different facial features are occluded.

In the Baseline group, where there are no occlusions in the faces within the images, the training set comprises 10,400 fake images and 10,400 real images. The validation set includes 1,300 fake images and 1,300 real images, and the test set similarly contains 1,300 fake and 1,300 real images. This results in a total of 26,000 images for the Baseline group.

The Both Eyes group, where all the training images feature occlusions in both eyes, mirrors the distribution of the Baseline group. It includes 10,400 fake and 10,400 real

images in the training set, 1,300 fake and 1,300 real images in the validation set, and 1,300 fake and 1,300 real images in the test set, totaling 26,000 images.

Similarly, the Right Eye group, in which all training images have occlusions in the right eye, includes 10,400 fake and 10,400 real images in the training set, 1,300 fake and 1,300 real images in the validation set, and 1,300 fake and 1,300 real images in the test set, resulting in a total of 26,000 images.

The Left Eye group, where all training images have occlusions in the left eye, maintains the same distribution, with 10,400 fake and 10,400 real images in the training set, 1,300 fake and 1,300 real images in the validation set, and 1,300 fake and 1,300 real images in the test set, summing to 26,000 images.

The Nose group, with occlusions in the noses of all training images, also includes 10,400 fake and 10,400 real images for training, 1,300 fake and 1,300 real images for validation, and 1,300 fake and 1,300 real images for testing, totaling 26,000 images.

For the Mouth group, where occlusions are applied to the mouths in all training images, the distribution remains consistent, with 10,400 fake and 10,400 real images for training, 1,300 fake and 1,300 real images for validation, and 1,300 fake and 1,300 real images for testing, making up 26,000 images.

The Jawline group, featuring occlusions in the jawlines of all training images, follows the same pattern, with 10,400 fake and 10,400 real images in the training set, 1,300 fake and 1,300 real images in the validation set, and 1,300 fake and 1,300 real images in the test set, totaling 26,000 images.

Similarly, the Forehead group, with occlusions in the foreheads of all training images, includes 10,400 fake and 10,400 real images for training, 1,300 fake and 1,300 real images for validation, and 1,300 fake and 1,300 real images for testing, making a total of 26,000 images.

Finally, the Chin group, where all training images have occlusions in the chin area, maintains the same distribution, with 10,400 fake and 10,400 real images for training, 1,300 fake and 1,300 real images for validation, and 1,300 fake and 1,300 real images for testing, totaling 26,000 images.

Overall, each training group in Phase 3 sustains a balanced and consistent number of images across the different sets, ensuring a fair and rigorous evaluation of the model's capability to detect manipulations across various facial regions. This structured methodology aids in understanding the impact of occluding different facial features on the model's performance in detecting deepfakes.

Dataset	Groups	Train		Validation		Test		Total
		Fake	Real	Fake	Real	Fake	Real	
Combined Dataset	Baseline	10400	10400	1300	1300	1300	1300	26.000
	Both eyes	10400	10400	1300	1300	1300	1300	26.000
	Right eye	10400	10400	1300	1300	1300	1300	26.000
	Lift eye	10400	10400	1300	1300	1300	1300	26.000
	Nose	10400	10400	1300	1300	1300	1300	26.000
	Mouth	10400	10400	1300	1300	1300	1300	26.000
	Jawline	10400	10400	1300	1300	1300	1300	26.000
	Forehead	10400	10400	1300	1300	1300	1300	26.000
	Chin	10400	10400	1300	1300	1300	1300	26.000

Table 3.3 Summary of Dataset Size for Each Group in Phase 3

3.4.2 Model Selection

For the deepfake detection algorithm, deep convolutional models were chosen as feature extractors: EfficientNet-B7 and XceptionNet. Both models were initialized with pre-trained ImageNet weights, enabling rich feature representations learned from a large-scale dataset of natural images to be leveraged.

In the intricate field of deepfake detection, the exploration of various architectural innovations has been paramount to augment both the accuracy and computational efficiency of predictive models. The XceptionNet architecture, eloquently introduced by François Chollet[161], revolutionized the conventional convolution operations by introducing depth-wise separable convolutions. This architectural nuance dissects a standard convolution operation into two discrete processes, namely, depth-wise and point-wise convolution operations, thereby strategically bifurcating the computational process.

This methodological division significantly alleviates the computational burden by reducing the number of parameters and computational resources requisite for the model. Such an approach not only engenders a hastened training process but also enhances the model's capacity for improved generalization performance, thereby mitigating the risk of overfitting despite the model's complexity. The efficacy of the XceptionNet architecture is not merely theoretical but has been empirically substantiated in the realm of deepfake detection, most notably by Rossler et al. [158]. Their research not only corroborates the architectural prowess of XceptionNet in accurately detecting

deepfakes but also underscores its utility in practical applications, thereby validating its adoption in various research contexts.

Further, the depth-wise separable convolutions utilized in XceptionNet minimize the redundancy observed in traditional convolution operations, thereby ensuring that each parameter is optimally utilized in capturing pertinent features in the input data. This optimal utilization of parameters is particularly pivotal in scenarios with limited computational resources, enabling researchers to deploy more complex models even in constrained environments. In light of its distinguished performance and empirical validations in deepfake detection, XceptionNet has been prominently featured as a critical feature extractor in the present investigative study. This underscores the model's potential not only as a standalone predictive model but also as a potent feature extractor in conjunction with other models, paving the way for hybrid approaches in tackling the multifaceted challenge of deepfake detection. Due to its excellent performance in deepfake detection, XceptionNet was used as a feature extractor in the present study.

Meanwhile, EfficientNet-B7 is the largest variant in the EfficientNet architecture family based on its depth and number of parameters, achieving the highest performance among all EfficientNet models, as shown in Figure Figure 3.4. It achieved state-of-the-art results on the ImageNet dataset, with a top-1 accuracy of approximately 84.3% and a top-5 accuracy of 97.0%. EfficientNet-B7 has a significantly smaller number of parameters compared to some leading CNN models, making it up to 8.4 times smaller in terms of parameter count, depending on the specific comparison. This reduction in parameters translates to fewer calculations, contributing to its faster processing speed. Specifically, EfficientNet-B7 can be up to 6.1 times faster in terms of inference speed compared to some leading CNN models, meaning it processes data and generates predictions more quickly. Additionally, this efficiency can also benefit training times, including the time to train each batch and potentially reducing the overall training duration [2].

Moreover, EfficientNet-B7 was pre-trained using a technique called Noisy Student, which involves adding noise to the data and training the model on the noisy data. This helps make the model more robust to variations in input data and improves its performance on downstream tasks. For this reason, EfficientNet-B7 from the EfficientNet family was chosen to be the second feature extractor in this research.

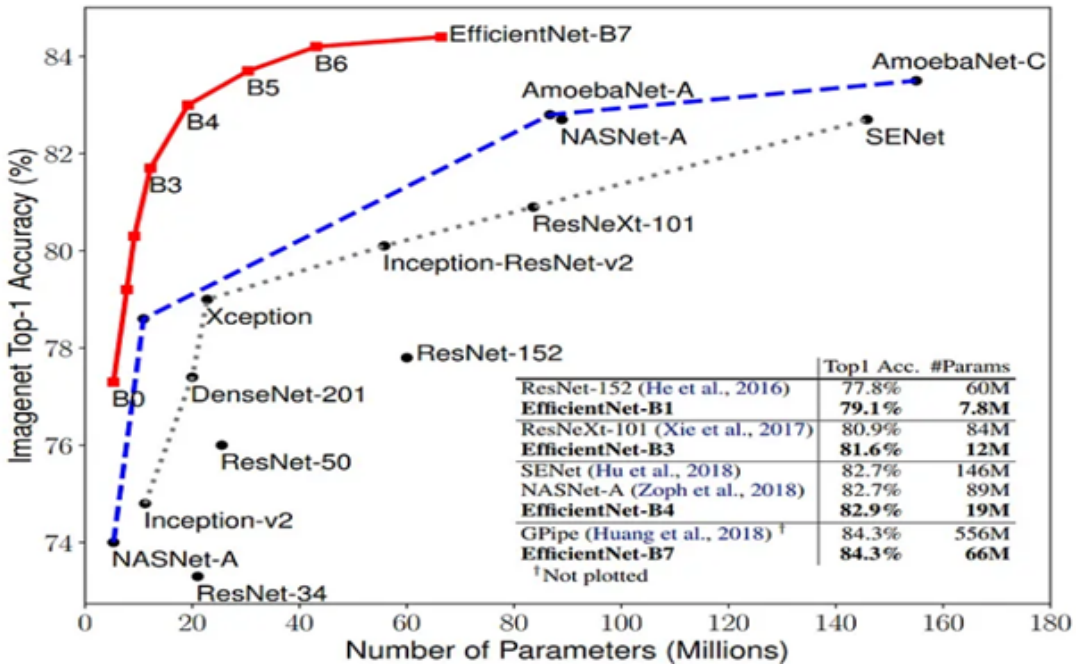


Figure 3.4 Performance and Size Comparison of EfficientNet Models on the ImageNet Dataset [2]

3.4.3 Pre-processing and Training Set-up

The dataset utilized in this research comprises a vast array of both authentic and digitally manipulated videos, all encoded in the MP4 format. We incorporated two popular datasets: Celeb-DF v2 [159], representing the first generation of deepfake datasets, and FaceForensics++ (FF++) [158], indicative of the second generation. This carefully curated collection aims to facilitate an in-depth exploration of video manipulation detection and analysis. To optimize the dataset for machine learning applications, we extracted frames from both the original and altered videos, selecting 16,000 frames from each category.

To focus on the critical visual data, especially facial features commonly targeted for manipulation, we employed the OpenCV library's capabilities. Using the faceCascade, a pre-trained face detection model, we cropped the facial regions from the extracted frames, effectively isolating them from the background. This selective cropping reduces computational demands and streamlines the AI model's learning process. We adjusted the scaleFactor to 1.3, shrinking the image size by 30% at each detection step to accommodate faces of various sizes. The minNeighbors parameter was set to 5, balancing detection sensitivity and accuracy. The detection results, stored as a list of

rectangles where each rectangle represents a detected face, were highlighted in green (0, 255, 0) with a thickness of 4 pixels. After manual verification, faces were cropped and saved into two folders: one containing real faces and the other fake faces.

To enhance our focus on facial features, we utilized MediaPipe's face mesh technology to accurately map facial landmarks, identifying key regions for manipulation detection. We generated two sets of face cut-outs:

- Cut-out 1: Targeting landmark positions for the chin, forehead, mouth, and jawline.
- Cut-out 2: Focusing on the left eye, right eye, both eyes, and nose.

The cut-outs were applied to the selected datasets, with pixels in the detected regions replaced by a value of 255 to create a white occlusion. For instance, to obscure the eye regions, we drew filled white circles. A thickness value of -1 indicates that the circle is to be completely filled, resulting in a solid disk rather than merely an outlined circle.

For organizational efficiency, processed images were systematically categorized and stored in folders labeled according to the specific facial region extracted—face cut-out 1 or face cut-out 2, as illustrated in Figure 3.2. This approach facilitated streamlined data access and manipulation.

To ensure dataset consistency and reliability, we applied a rigorous standardization process. This involved image normalization techniques, setting a uniform per-channel mean of (0.485, 0.456, 0.406) and a standard deviation of (0.229, 0.224, 0.225). Isotropic resizing achieved a standardized resolution of 224 x 224 pixels for all images, with zero padding as necessary to preserve aspect ratios. Image augmentation techniques, including Image Compression, Gaussian Noise, and Flipping, were introduced with a 10-15% probability during training but not during testing or validation.

Model optimization was conducted using the Rectified Adam optimizer, with an initial learning rate of 0.001 and a weight decay of 0.0005. The learning rate was dynamically adjusted using a Reduction on Plateau strategy, reducing the rate by 0.25 after 2 consecutive epochs without improvement. The training was governed by the Binary Cross-entropy Loss function and limited to 30 epochs, with an early stopping mechanism to prevent overfitting.

The experimental framework was structured into three phases, each designed to examine different aspects of facial manipulation detection. Experiments were conducted

using a high-performance NVIDIA Geforce RTX 3080 Ti Laptop GPU and Google Colab Pro+. The first phase applied the Cut-out Technique to both datasets, preparing three sets of images per dataset: a baseline set with no facial augmentation, and two sets with specific face cut-outs. The experimental procedure was structured into three distinct phases.

- **Phase One: Apply the Cut-out Technique with Each Dataset**

During this phase, for each dataset (FaceForensics++ and Celeb-DF v2), three distinct sets of images were separately prepared in the preprocessing stage. These sets included: baseline images with no facial alterations; Cut-out 1, featuring images with cut-outs in four different regions - the chin, mouth, jawline, and forehead; and Cut-out 2, consisting of images with cut-outs in four separate regions - the left eye, right eye, both eyes, and the nose. Following this, both deep learning models were trained using these three distinct sets. The models' performance was subsequently evaluated on datasets that they had not been exposed to during the training phase.

This phase aimed to establish a foundational understanding of how different types of facial occlusions impact the model's ability to detect manipulations. We sought to determine which facial features carry the most information for face detection by comparing core features (such as eyes and nose) versus external features (such as jawline and forehead). Additionally, we assessed whether the AI model's performance improves by concentrating on the same facial features used in the medical field for training individuals with prosopagnosia . By examining the model's response to these specific occlusions, we aimed to gather insights into the importance of different facial regions in manipulation detection and to draw parallels to existing practices in both medical diagnostics and standard facial recognition technology.

- **Phase Two: Apply the Cut-out Technique with the Combined Dataset**

In phase 2, to enhance the model's generalization capabilities, we combined the two datasets, FaceForensics++ (FF++) and Celeb-DF v2, and trained the model to observe its performance across a more diverse dataset. As in phase one, we utilized three distinct training groups: baseline (with no facial alterations), Cut-out 1 (images with cut-outs in the chin, mouth, jawline, and forehead), and Cut-out 2 (images with cut-outs in the left eye, right eye, both eyes, and nose). Both deep learning models were then trained using these three groups, and their performance was subsequently

evaluated on previously unseen datasets to assess their robustness and generalization to new data.

- **Phase Three: Apply the Cut-out Technique for Each Facial Feature**

In phase 3, we conducted targeted tests on each facial region to better understand their individual roles in deepfake detection. During this phase, the model was independently trained on images where specific facial regions, such as the eyes, were occluded. This approach was systematically replicated for each facial region, resulting in a total of nine training sets, including the baseline.

Each set focused on a different facial area, and the model's performance in distinguishing between real and fake faces was assessed for each set. For evaluation, we used a set of validation images that were different from those in the training dataset. This phase zoomed in on the role of specific facial features in the AI models' ability to detect deepfakes, providing detailed insights into the contribution of each facial region to the overall detection accuracy.

3.4.4 Testing the Models

Each model in this study was rigorously evaluated using the original, non-augmented dataset, with a specific allocation of 10% of this dataset set aside exclusively for testing purposes. This allocation strategy is crucial to ensure a robust evaluation of the model's performance. It is imperative to emphasize that the test dataset must be entirely distinct from the data used during the training and validation phases. This separation is essential to guarantee that the model's performance is assessed on completely new and unseen data, thereby providing a true measure of its effectiveness and generalizability in real-world scenarios.

For the purposes of testing, only the face regions within the images were utilized. This focused approach was adopted with the objective of specifically evaluating the model's proficiency in detecting fake faces by analyzing facial features. By concentrating on the facial regions, the study aims to ascertain the model's ability to discern subtle discrepancies and anomalies that are characteristic of deepfakes. This targeted evaluation is critical in understanding the model's capabilities in the context of facial recognition and deepfake detection, where the accuracy of identifying and differentiating facial features plays a pivotal role. The results of this testing phase are expected to provide valuable insights into the effectiveness of each model in detecting deepfakes,

thereby contributing to the advancement of reliable and efficient deepfake detection technologies.

3.4.5 Performance Measurement

We decided to use the accuracy ,the area-under-curve (AUC) of ROC, and log-loss would be used to measure model performance. The AUC score summarizes the relationship between the false positive rate (FPR) and true positive rate (TPR) of the binary classifier, and is a widely used metric in machine learning and statistical analysis to evaluate the performance of classification models by measuring the degree to which it is capable of distinguishing between positive and negative classes. In a classification problem, the AUC is a measure of the probability that a randomly selected positive sample will be ranked higher than a randomly selected negative sample. With values ranging from 0 to 1, 0 means that the model performs no better than random guessing while a value of 1 indicates perfect classification.

Log-loss is also known as logarithmic loss or cross-entropy loss, and is another metric commonly used in evaluating the performance of classification models. It measures the difference between the predicted probability distribution and the true probability distribution for a set of samples. Log-loss is defined as:

$$\text{LogLoss} = -\frac{1}{N}\sum_{i=1}^{N}[y(i)\log(p(i)) + (1-y(i))\log(1-p(i))] \tag{3.1}$$

Where:
N is the number of samples in the dataset.
y(i) is the actual label for sample i, where 0 indicates the sample is fake, and 1 indicates the sample is real.
p(i) is the predicted probability of sample i being positive (being fake in this context).

The log-loss metric penalizes models more heavily for incorrect predictions with high confidence, where the predicted probability is close to 0 or 1, compared to predictions where the confidence level is lower (that is, closer to 0.5). This is because the logarithmic function amplifies the differences between predicted and true probabilities for extreme values.

In the context of deepfake detection, log-loss can also be used to rank different models based on their ability to distinguish between real and fake images. The lower

the log-loss value, the better the model is at predicting the true labels of the samples in the dataset.

3.4.6 Relevant libraries and Toolkits

In this research endeavor, a variety of Python libraries were utilized for diverse computational tasks. Initially, the VideoCapture function from the OpenCV library facilitated the extraction of frames from video files. Additionally, the ImageDataGenerator library was employed for the augmentation of images and preparation of data for the training process.

Other integral Python libraries, including Matplotlib, NumPy, Pandas, and random, were incorporated into the main script for functions such as generating visualizations, data manipulation, and randomization. For the specific task of face detection and cropping, the MediaPipe libraries, coupled with the Face Mesh , were utilized, leveraging facial landmarks for precise face cut-outs.

Moreover, scikit-learn, a comprehensive and open-source machine learning library for Python, played a pivotal role in the testing phase of the models. This library, adept in both supervised and unsupervised machine learning, provided an extensive selection of algorithms for classification, regression, clustering, and dimensionality reduction. Its construction upon well-known libraries like NumPy and SciPy facilitated seamless integration into our project framework.

In our study, scikit-learn was paramount in importing crucial evaluation metrics, such as log loss, accuracy, and the Area Under the Curve (AUC). These metrics were instrumental in assessing the performance of the trained models and pinpointing potential areas for enhancement. Scikit-learn proved especially valuable for classification tasks, which are central to our study's objective of categorizing input into specific categories and subsequently generating accurate output values.

3.5 Chapter Summary

In this chapter, a systematic and comprehensive methodology encompassing seven distinct stages is proposed for the identification of pivotal facial features in deepfake detection. This methodology integrates renowned datasets, such as FaceForensics++ and Celeb-DF, alongside state-of-the-art deep convolutional neural network models,

including EfficientNet-B7 and XceptionNet, complemented by a specialized face cut-out technique to accentuate the most informative features.

The systematic nature of this methodology underpins its robustness and precision, hallmarks of rigorous academic inquiry. Additionally, it offers a structured framework for ongoing research in the realm of deepfake detection, allowing other researchers to adopt or adapt this methodology in accordance with their unique research objectives.

We posit that our methodological approach will contribute significantly to the field of deepfake detection. It aims not only to pinpoint essential facial features but also to facilitate the development of increasingly efficient and effective techniques for deepfake identification.

Chapter 4

Results and Evaluation

4.1 Introduction

In this chapter, a comprehensive evaluation of the performance of the proposed method is conducted for various contexts, and the visual outcomes are examined which result from the application of the face cut-out augmentations with the Forensics++ and Celeb-DF datasets. Additionally, a comparative analysis is conducted of the results obtained when the models were trained with these datasets in three distinct settings.

- Baseline: The original faces without any augmentation.
- Cut-out 1: Four cut-outs strategically applied to the chin, mouth, jawline, and forehead regions of the face.
- Cut-out 2: Four cut-outs strategically applied to the left eye, right eye, both eyes, and nose regions.

Moreover, this study includes a comprehensive comparison between the results obtained from our models and those achieved using current state-of-the-art deepfakes detection techniques. This comparative analysis is conducted under identical conditions, utilizing the same datasets and methodologies to ensure a fair and accurate assessment. By aligning the experimental setup across different models, the study aims to provide a clear, objective comparison of the effectiveness of various approaches in deepfakes detection. This comparison is crucial in understanding how the proposed models stand in relation to existing technologies and in identifying any areas where they may offer improvements or present new challenges.

The results obtained from each phase of the experiment are methodically presented, beginning with the initial outcomes from the preliminary testing phase and progressing through to the more advanced stages of the evaluation. This sequential presentation of results allows for a clear and logical understanding of the models' performance over time and under different conditions. It also facilitates a detailed analysis of the models' strengths and weaknesses, providing insights into their efficacy in various scenarios.

4.2 Phase One: Evaluation of the Performance of the Cut-out Technique with Each Dataset Individually

During this phase, three image groups were generated from each dataset employed in the study: Baseline, Cut-out 1, and Cut-out 2. Subsequently, these groups were trained using the two deep convolutional models selected, which are EfficientNet-B7 and XceptionNet.

The obtained results elucidate that models trained using the face Cut-out 2 group exhibited superior performance compared to those trained with the Baseline and face Cut-out 1 groups. This significant improvement is depicted in Figure 4.1, which illustrates the comparative results of the three groups when trained with both EfficientNet-B7 and XceptionNet models across the FF++ and Celeb datasets. Furthermore, it was observed that the performance of the Cut-out 1 group was inferior to the Baseline group in certain instances. This is particularly evident in the outcomes derived from training with the EfficientNet-B7 model, where the Cut-out 1 group's performance lagged behind that of the Baseline group, as detailed in Table 4.1. This table presents the results from Phase 1, comparing the three groups—Cut-out 1, Cut-out 2, and Baseline—across each dataset (FF++ and Celeb) with both models (EfficientNet-B7 and XceptionNet), utilizing three key performance metrics: accuracy, AUC (Area Under the Curve), and log-loss.

As delineated in Table 4.1, the data elucidates substantial performance improvements in both EfficientNet and Xception models consequent to the adoption of Cut-out 2. Specifically, the AUC (Area Under the Curve) percentage shows an improvement ranging from 3.7% to 6.9% over the Baseline metrics in the EfficientNet-B7 model. For the Xception model, the AUC improvements from Cut-out 1 to Cut-out 2 are more modest, ranging from 1.1% to 2.47%. This highlights the effectiveness of Cut-out 2 in

Figure 4.1 Phase 1 Test Results Overview

enhancing model performance, particularly in the EfficientNet-B7 model. In terms of accuracy, the comparison between Cut-out 2 and Cut-out 1 within the EfficientNet-B7 model demonstrates notable performance gains, with improvements ranging from 2.22% to 21.2%. In contrast, for the Xception model, the differences in accuracy between Cut-out 2 and Cut-out 1 are minimal, indicating comparable levels of performance across both groups.

These findings underscore the efficacy of the Cut-out 2 technique in training facial recognition models, particularly in the EfficientNet-B7 model. This can be attributed to the enforced learning within the Cut-out 2 dataset, where models are compelled to discern faces by focusing on critical facial regions—regions that provide essential cues in distinguishing fake faces from genuine ones.

In the context of the Celeb-DF dataset, the log-loss results for the Cut-out 2 group improved by 0.26 when compared to the Baseline group in the EfficientNet-B7 model. The Xception model exhibited a similar trend, with a log-loss improvement of 0.51. These results further confirm that models trained with face Cut-out 2 augmentations outperformed those trained with Cut-out 1 and baseline datasets.

When training the deep neural models on face Cut-out 2 images, the models prioritized the regions of the face that were exposed. For images in the Cut-out 2 group, the facial regions covered up were the left eye, right eye, both eyes, or nose.

Model/Training Group	FF++			Celeb-DF		
	ACC	AUC	logloss	ACC	AUC	logloss
EfficientNet-B7+ Baseline	0.77	0.81	0.59	0.91	0.89	0.51
EfficientNet-B7 + Cut-out 1	0.66	0.78	1.12	0.90	0.88	0.44
EfficientNet-B7+ Cut-out2	0.80	0.84	0.53	0.92	0.93	0.25
Xception+ Baseline	0.77	0.77	0.78	0.84	0.79	0.80
Xception + Cut-out 1	0.75	0.81	0.90	0.90	0.91	0.35
Xception+ Cut-out 2	0.75	0.83	0.76	0.91	0.92	0.29

Table 4.1 Phase one Test Results Overview

The results suggest that the regions of the face outside of these areas provide more significant information in discerning disparities between authentic and synthetic faces.

Consequently, it is concluded that the superior performance of the Cut-out 2 augmentations is due to the inclusion of facial features other than the eyes and nose, which are central features of the face. This allows for more reliable detection of facial dissimilarities. Huang et al. [162] investigated expression recognition under conditions where parts of the face were occluded. Their findings revealed that models could identify a majority of facial expressions even when the eyes were occluded by leveraging the external features of the face for cues. This observation aligns with the findings of the present study, which indicate that training models with datasets incorporating face cutouts enhances the model's ability to recognize deepfake images more effectively than training with baseline datasets. Specifically, when comparing the two groups with cutouts, the models demonstrated superior performance with the Cut-out 2 group (comprising cutouts in regions such as both eyes, left eye, right eye, and nose) compared to the Cut-out 1 group (comprising cutouts in regions such as the jawline, mouth, forehead, and chin).

4.3 Phase Two: Evaluation of the Performance of Cut-out Technique with the Combined Dataset

In this phase of the study, the datasets utilized in phase one (FF++ and Celeb-DF) were combined to increase the overall volume of training data. This integration was expected to enhance the model's generalization capabilities and performance on previously unseen data. The enriched dataset, containing a more diverse array of

examples, aimed to facilitate the models in better identifying and learning underlying patterns and features within the data.

Table 4.2 presents the results from the second phase of the experiment, where the efficacy of three distinct groups—Baseline, Cut-out 1, and Cut-out 2—was assessed using the combined dataset of facial images. During this phase, the models were trained for 30 epochs, and their performance was evaluated using several metrics, including accuracy (ACC), area under the curve (AUC), and log-loss. The results indicate that the EfficientNet-B7 model, when trained with Cut-out 2, achieved the best performance, with an AUC of 0.91, accuracy of 0.89, and a log-loss of 0.45. Similarly, the Xception model trained with Cut-out 2 demonstrated improved results, reaching an AUC of 0.88, accuracy of 0.86, and a log-loss of 0.85. In contrast, the baseline datasets for both the EfficientNet-B7 and Xception models showed lower performance, with AUC values of 0.87 and 0.73, respectively.

These results suggest that implementing Cut-out 2 significantly enhances the efficacy of face recognition models, outperforming both Cut-out 1 and baseline data. Additionally, the EfficientNet-B7 model demonstrated stronger performance compared to the Xception model in this phase. The nature of the cutouts in the Cut-out 2 group appears to be instrumental in preserving critical facial features, helping the model more accurately learn these features and thereby improving its performance in deepfake detection.

Moreover, the AI models exhibited improved proficiency in identifying deepfakes when analyzing external facial regions such as the forehead, cheeks, and chin. This observation aligns with findings from other studies (e.g., [163, 156]), which suggest that peripheral facial regions may contain crucial information for distinguishing between genuine and fabricated faces, a key aspect in digital authenticity verification.

Interestingly, the findings run counter to expectations from prior research on prosopagnosia, which suggests that focusing on central facial features like the eyes and nose improves face recognition. Despite this, the models in this study performed better when trained on datasets where the eyes and nose were occluded (as in the Cut-out 2 group), compared to datasets where external regions like the jawline, mouth, forehead, and chin were occluded (as in the Cut-out 1 group). Additionally, it should be noted that the test dataset contained fully visible facial features, with no cutouts applied.

Several potential explanations could account for these findings. One possibility is that external facial regions may offer more consistent information, as they are less affected by variations in lighting, pose, or expression compared to central features.

Model/Training Group	Combined dataset		
	ACC	AUC	logloss
EfficientNet-B7+ Baseline	0.90	0.87	0.69
EfficientNet-B7 + Cut-out 1	0.89	0.90	0.48
EfficientNet-B7+ Cut-out2	0.89	0.91	0.45
Xception+ Baseline	0.77	0.73	0.94
Xception + Cut-out 1	0.83	0.85	0.84
Xception+ Cut-out 2	0.86	0.88	0.85

Table 4.2 Phase Two Test Results Overview

Consequently, these regions might retain more stable and reliable information, which can be instrumental in differentiating real and fake faces.

These findings are supported by a study by Nirkin et al. [163], which examined the manipulation of facial regions and found that deepfake manipulations are often concentrated in central facial features, leaving peripheral regions relatively unchanged. This observation was particularly evident in the FF++ datasets used in this research, where manipulations were typically confined to a square area at the center of the face, with the surrounding regions left untouched. This could explain why external facial features emerged as more informative in this study, offering critical insights for improving deepfake detection methods.

Overall, the findings suggest that external facial regions may be as important, or even more important, than central facial features in detecting deepfakes. To gain a more nuanced understanding of how each region of the face contributes to deepfake detection, further tests were conducted on each facial region, which are discussed in the next section.

4.4 Grad-CAM Visualization

Based on the results shown in tables 4.1 and 4.2 from Phase 1 and Phase 2, we observed that the accuracy of the Cut-out 2 approach did not consistently improve model performance. The face-out method, specifically, does not work better in some cases and performs better in others. For instance, when we trained the Xception model using the FF++ dataset, the accuracy of the Cut-out 2 group did not surpass that of

the Cut-out 1 group, with both achieving an accuracy of 75%. In contrast, the Baseline group showed a better result with an accuracy of 77%.

To gain further insight, we utilized another metric, the Area Under the Curve (AUC). The AUC results indicated that the Cut-out 2 group achieved a higher score compared to both the Cut-out 1 and Baseline groups. This finding was also supported by the results using the log loss metric. These results suggest that accuracy alone may not be a clear indicator of model performance in classification tasks.Another possible reason could be the nature of the FF++ dataset, which employs four different techniques to create fake videos, potentially influencing the model's ability to learn effectively.

Another notable observation regarding model performance accuracy was found in Phase 2, where a similar outcome occurred with the EfficientNet model. Both the Cut-out 1 and Cut-out 2 groups achieved an accuracy of 89%, which was equal in performance. Given that in Phase 2 we combined the FF++ and Celeb-DF v2 datasets, it is plausible that the neutrality of the FF++ dataset and the mixed dataset in Phase 2 influenced these results. Alternatively, it may indicate that accuracy is not the most appropriate metric to assess performance in these cases.

To further investigate the model's decision-making process in classifying images as real or fake, we will utilize Gradient-weighted Class Activation Mapping (Grad-CAM) [164]. Grad-CAM offers a method to interpret the decision-making of Convolutional Neural Network (CNN)-based models. By visualizing the effects and impacts of our augmentation techniques on training, we can ascertain whether these augmentations assist the models in more accurately identifying fake regions rather than simply memorizing facial features.

We investigated the Class Activation Map (CAM) output of the Xception model trained on the FF++ dataset. Notably, models trained with both face cutout groups (cutout 1 and cutout 2) focused on similar central facial regions for fake image detection. Specifically, the face cutout 2 group emphasized areas encompassing the upper lip, nose, and extending to the right eye, whereas the face cutout 1 group focused on the same areas but extended to the left eye instead of the right eye as shown in Figure 4.2. Figure 4.2 illustrates real and fake images, the face mask for the fake pixels in the test image, and the CAM output for the model, including the Grad-CAM output for the baseline, Face-Cutout 1, and Face-Cutout 2 groups.

Both cutout groups achieved comparable accuracy (75%, Table 4.1). However, the baseline group, consisting of images without face cutouts, achieved a higher accuracy

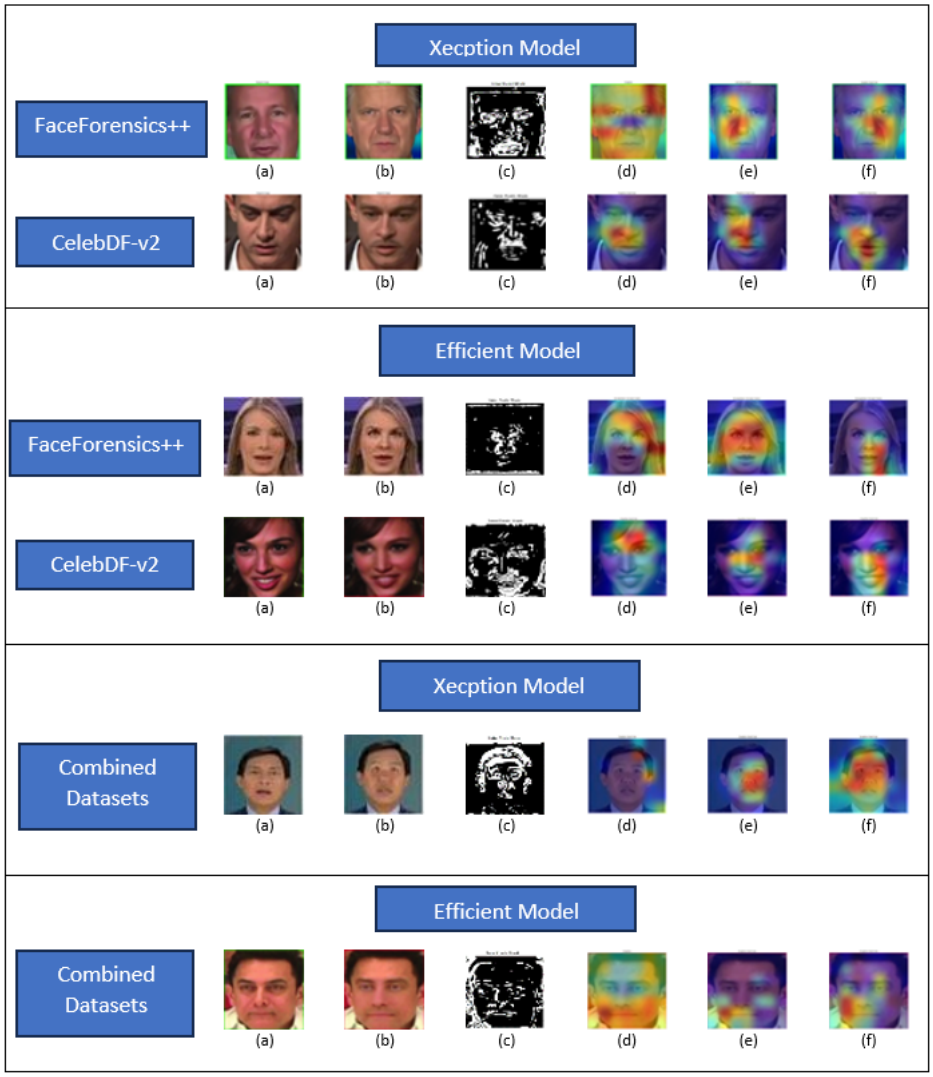


Figure 4.2 Face extracted from frames of both a real and a corresponding fake : (a) Real face, (b) DeepFake, (c) SSIM difference mask revealing manipulated pixels, (d) GradCAM output from a baseline model, (e) GradCAM output from a Face-Cutout 1 trained model, and (f) GradCAM output from a Face-Cutout 2 trained model.

(77%, Table 4.1). Analysis of the baseline model's CAM output revealed activation across the entire face, including potentially irrelevant areas, suggesting overfitting. This overfitting likely contributes to the higher, albeit potentially misleading, accuracy.

Next, we explored the Xception model with a different dataset (CelebA). Here, the model trained with face cutout 2 group achieved superior accuracy (91%) compared to other groups. The CAM output for this model indicated a focus on the nostrils, mouth, and lower left face region, including the jawline and chin. Conversely, the cutout 1 model, with a lower accuracy, focused on the upper lip, nose, and left eye area, which also coincides with manipul ated regions in the CelebA dataset. These observations suggest that the mouth region plays a critical role in distinguishing deepfakes within the CelebA dataset. This aligns with our findings from phase three, where individual region training indicated the mouth held greater significance than the nose, and the eyes were more crucial than other facial areas.

The Efficient model demonstrated superior accuracy when trained with the FF++ dataset using the "face cutout 2" group. This group consisted of image patches positioned in the central regions of the face, specifically around the right eye, left eye, both eyes, and nose. To better understand the model's accuracy at this stage, we analyzed the Grad-CAM output to identify the image regions the model emphasized in its decision-making process. As illustrated in the accompanying Figure 4.2, the model focused on the area started from the lower right side of the face covring the Jawline ,the mouth ,chine , nose, and right eye most of these regiones identified as containing fake elements based on the fake pixel mask image for the test image. This finding suggests that training the model with the "face cutout 2" group enhances its ability to detect fake images, which explains the higher accuracy of 80% achieved with this group. In contrast, the "face cutout 1" group included these regions but also encompassed other parts such as the cheeks and hair, resulting in a lower accuracy of 66%.

The Efficient model exhibited optimal performance when trained with the Celeb dataset. Specifically, the model trained with the "face cutout 2" group derived from the Celeb dataset achieved an accuracy of 92%. The Grad-CAM output for this group demonstrates that the model effectively identified the fake regions of the face, as shown in Figure 4.2. This indicates that the model utilized these fake regions to accurately classify the images as fake, demonstrating effective learning during training without merely overfitting to the training data.

In analyzing the results from combining the FF++ and Celeb datasets, we trained the Efficient model with three groups: face cutout 1, face cutout 2, and the baseline.

The objective was to determine if the model could generalize better by training with diverse data types and a larger dataset. The model achieved an accuracy of 89% for both face cutout groups. To understand this similarity in accuracy, we examined the Grad-CAM output, which showed that the model focused on nearly the same facial areas for decision-making, as shown in the figure. These regions were identified as fake based on the mask showing fake pixels in the test images, indicating that the model benefited from training with the face cutout groups. For the baseline group, the model appeared to overfit, highlighting the entire image, which explains the slightly higher accuracy of 90% achieved in this scenario.

In the analysis of the Xception model with combined datasets, the "face cutout 2" group performed the best. The Grad-CAM output indicated that the model concentrated on the fake regions of the image, such as the center of the face, forehead, and areas near the hairline, to make its decisions. However, the model seemed to overfit with the baseline group by focusing on background of the images parts of the images that did not include any facial regions.

The results of our study indicate that the Efficient model achieves higher accuracy in detecting fake images compared to other models, particularly when trained with the Celeb dataset. Training with the "face cutout 2" group, which consisted of image patches positioned in the central regions of the face (specifically around the right eye, left eye, both eyes, and nose), significantly enhances the model's detection capabilities. This is evidenced by the 80% accuracy achieved with the FF++ dataset and the 92% accuracy with the Celeb dataset. Grad-CAM analysis further supports these findings, showing that the model effectively targets regions containing fake elements, thereby validating its decision-making process.

Our observations from the Grad-CAM output for the face cutout 2 group revealed that it achieved higher accuracy by focusing on the mouth and lower side of the face when distinguishing fake images. This targeted approach allowed the model to make more accurate decisions.

In contrast, the face cutout 1 group achieved the second-highest accuracy after the cutout 2 group. Based on CAM output, there are similarities in the regions used by the face cutout 2 group; however, the face cutout 1 group focused more on the center to upper left side of the face. The inclusion of additional genuine parts likely resulted in lower accuracy.

Interestingly, there were two instances where both the cutout 1 and cutout 2 groups achieved the same accuracy: once with training the Xception model with the FF++

dataset and again with training the Efficient model with combined datasets. CAM analysis for these two groups showed that both focused on the same regions of the face to make decisions, which are the regions indicating fake pixels in the images. This explains the similar accuracy for both groups, as both effectively identified the correct fake parts of the images.

In some cases, the baseline group outperformed the face cutout groups. However, CAM analysis for these cases revealed signs of overfitting. The baseline group, trained on the entire image, exhibited signs of overfitting, resulting in a marginally higher but potentially misleading accuracy . This suggests that the baseline model fixated on irrelevant details within the training data, hindering its generalizability to unseen fake images.

Overall, these findings underscore the importance of using targeted face cutouts and diverse datasets to improve the accuracy and generalization capabilities of models in detecting fake images. Several factors influence model performance in detecting deepfakes, including the quality and variety of training data and the techniques used to create the fakes. A significant source of error arises from the lack of robust features, particularly closed or partially closed eyes and mouths. Addressing these challenges is crucial for developing more reliable and accurate deepfake detection models.

4.5 Phase Three: Evaluation of the Cut-out Technique for Each Facial Feature

In an effort to deepen our understanding of which facial features are crucial for detecting deepfake images, we embarked on an in-depth study. This study involved applying cut-out techniques to isolate specific facial areas, thereby allowing for the independent training and assessment of each individual region. We created several training groups based on each face cut applied, which informed the training of the models during phases one and two. Consequently, we generated 8 groups, in addition to the baseline group, each focusing on images featuring a single facial cut-out, such as the eyes or mouth. After establishing these groups, we trained them using two models, Xception and EfficientNet, to conduct a comprehensive evaluation of the role of each face cut. This evaluation is depicted in Table 4.3 that presents the accuracy of training the models with each group.

Figure 4.3 Performance Analysis of Deep Learning Models Trained Independently with Phase Three-Generated Datasets

A notable observation from the study is the variance in accuracy based on the balance between images with no cuts and those with specific facial regions obscured. The accuracy tends to improve when the AI models are not exposed to the entire face. We then compared the accuracy to determine which part of the face, when obscured, gives the models a higher accuracy.

It is important to note that the size of the training dataset used for all models is the same, ensuring a fair comparison. The goal was to assess how effectively each group could differentiate between synthetic (deepfake) and real images. This approach aimed to illuminate which facial regions are most valuable in identifying deepfakes. The results, summarized in the following findings, are notable.

- EfficientNet-B7 demonstrates a lower log-loss score compared to Xception for almost all facial features, with the exception of the left eye. This indicates that EfficientNet-B7 generally outperforms Xception in classifying facial features, particularly for the nose, mouth, and jawline, where the performance difference is most marked.
- The baseline log-loss score is notably high, suggesting that models without any facial cut-outs perform poorly in identifying deepfakes. However, incorporating cut-outs of facial features such as the eyes, nose, mouth, jawline, forehead, and chin significantly enhances the accuracy of facial feature classification.

- The log-loss scores for cut-outs of the right and left eyes are similar, implying that both eyes are equally important in deepfake classification.
- The nose cut-out emerges as the most critical feature for image classification, evidenced by its substantially lower log-loss score compared to the eyes. This lower score indicates a higher likelihood of correct image classification by the model. The effective performance of the model without the nose information suggests that the nose region provides relatively less useful information for deepfake identification than the eyes.
- The mouth cut-out is nearly as crucial as the nose in deepfake classification, as indicated by their similar log-loss scores. These findings suggest that, in comparison to the eyes, the nose and mouth regions offer less useful information for deepfake classification.
- The jawline cut-out is slightly less important than the mouth for facial feature classification, as reflected by its marginally higher log-loss score.
- Among all facial features analyzed for deepfake classification, the forehead, both eyes, and chin exhibit the highest log-loss values as shown in Figure 4.3. This implies that the model's ability to classify deepfakes is significantly reduced when these regions are obscured, underscoring the importance of the forehead, eyes, and chin in containing critical information for deepfake classification.

These findings are in line with those of Das et al.[156], who also trained the EfficientNet-B4 and Xception models with three groups of images with the facial features of either the eyes, nose, or mouth obscured. They found that the model trained on images with obscured noses and mouths performed best, while the model trained on images with obscured eyes performed worst. These consistent outcomes between the present study and the findings of Das et al.[156] strengthen the validity and reliability of the observed trends concerning the impact of facial region cut-outs on model performance in classifying deepfakes.

Overall, this Phase highlights the significant impact of specific facial regions on the performance of deepfake detection models. The EfficientNet-B7 model consistently outperforms Xception, particularly when certain facial regions are obscured. The results indicate that obscuring the nose and mouth regions leads to the most accurate deepfake classification, while the eyes, forehead, and chin provide critical information for the models. Ensuring a fair comparison by using the same training dataset size for

Datasets	Accuracy	
	Xception	EfficientNet
Baseline	73	81
Both eyes	84	85
Right eye	87	86
Lift eye	83	87
Nose	89	91
Mouth	86	88
Jawline	86	87
Forehead	78	83
Chin	81	86

Table 4.3 Accuracy of Xception and EfficientNet Models on Isolated Facial Feature Groups

all models, the findings of this study contribute to a deeper understanding of which facial features are essential for improving the robustness and accuracy of deepfake detection systems.

4.6 Comparison with State-of-the-Art Methods

Given the markedly better performance exhibited by the face Cut-out 2 group when compared to the Cut-out 1 and Baseline groups, a thorough comparative analysis was conducted to compared the approach used in the present study with state-of-the-art methodologies that have employed identical datasets and techniques. This evaluation aimed to assess the effectiveness and robustness of the proposed technique in relation to existing approaches within the same experimental setting. The aim of this rigorous comparative analysis was to establish the competitiveness and potential advantages of the proposed method in the realm of deepfake detection.

4.6.1 Deepfake Detection Approaches Using the FF++ Dataset

Compared to other methodologies, the models in this study were trained with among the smaller sample sizes, as shown in Table 4.4. For instance, Rossler et al.[158] conducted training on a dataset consisting of approximately 388,000 images, while Khan and Dang-Nguyen [153] employed around 200,000 images when training their

models. This divergence in the size of the training dataset may help explain the superior accuracy attained by these studies in comparison to the approach used here.

Another factor influencing the outcome is the approach used to select facial landmarks and to conduct the cut-out procedure. In both of the aforementioned studies, landmarks to be covered in specific facial regions in the images were selected at random, resulting in the model being exposed to all areas of the face within the same group of data. Consequently, there were no constraints preventing certain parts of the face from appearing in specific subsets of the dataset, making it more difficult to determine which facial components exerted more influence than others.

In contrast, the methodology employed in the current study adheres to specific guidelines throughout the face cut-out procedure. For instance, in the face Cut-Out 2 group, the exclusion of landmarks focused on specific regions: the left eye, right eye, both eyes, or nose. This purposeful selection of regions allowed the models to give preference to facial features and focus their attention on information extracted from the external facets of the face.

Das et al.[156] avoided using random cut-outs of the face and instead selected specific regions to be covered in each group. Their study employed the removal of data in two groups: the 'sensory' and 'convex-hull' groups. The former group achieved the best performance and covered the eyes, nose, and mouth regions. In fact, this group outperformed the face Cutout 2 group in the present study. The reason for this may be that Das et al.'s sensory group covered three facial regions, two of which (the nose and mouth) were found to be less important for deepfake detection compared to other regions of the face, as shown in Table 4.4.

Moreover, the sensory group utilized both of these less important regions, which allowed the model to benefit from the regions of the face that provided more information than others. In contrast, the face Cut-out 2 group used in this study involved four cutouts, three of which were in the eye regions and the other in the nose region.

Nevertheless, our study exhibited superior performance compared to that of Lee et al.[155] for the face patch group. Their research achieved an accuracy rate of 72.79%. In contrast, our approach involved covering the eyes and nose and, notably, we worked with a smaller dataset than that employed in their study. In addition, the model in the present study outperformed those in other studies such as Afchar et al.[166] and Zhang et al.[60] which utilized the FF++ dataset for deepfake detection. Meanwhile, although Zhang et al. achieved an accuracy of 79.09% by randomly dropping out parts of the frames, in the present study an accuracy level of 84% was achieved. This may be

Approach	Accuracy	AUC	Number of datasets	Year
Rossler et al.[158]	90.60	N/A	388K	2019
Matern et al.[165]	82	N/A	5330	2019
Khan and Dang-Nguyen [153]	95.57	N/A	200K	2022
Lee et al. [155]Face-patch	72.79	N/A	60K	2022
Lee et al. [155] Face-crop	80.56	N/A	60K	2022
Afchar et al. [166]	83.10	N/A	16K	2018
Das et al.[156]	N/A	96.73	N/A	2021
Zhang et al. [60]	79.09	72.22	6706	2021
Lin et al.[167]	N/A	71.41	200K	2023
Our face Cut-out 2	84	80	32K	2024

Table 4.4 Comparative Performance Analysis of Baseline Models in Deepfake Detection on the FaceForensics++ Dataset.

said to confirm the hypothesis that the identification of facial differences is facilitated by the utilization of facial features that are not exclusively located at the centre of the face.

The novel facial Cut-out 2 could reduce overfitting and enhance the model's detection capabilities. Additionally, it is also demonstrated that the model used in the present study can learn from a smaller volume of data.

4.6.2 Deepfake Detection Approaches Using the Celeb-DF Dataset

The Celeb dataset is a widely used benchmark dataset used in the evaluation of the performance of deepfake detection models, since it contains a large number of high-quality videos of celebrities that have been manipulated to create deepfakes. By comparing the performance of various baseline models with this dataset, insights can be gained into which models are most effective in detecting deepfakes and how they compare with each other in terms of accuracy. Table 4.5 presents a comparison of different deepfake detection baseline models used with the Celeb dataset, highlighting the best results achieved by each model.

It can be observed that the approach proposed in this study achieves an impressive level of performance which is higher than that of all of the state-of-the-art models, with an accuracy of 92%, and AUC of 93%, thus demonstrating the effectiveness and ability of this approach in handling various deepfake generation methods.

Approaches	Accuracy	AUC	Year
Li et al.[168]	80.58	0.84	2020
Masi et al [169]	76.6	0.82	2020
Haliassos, et al.[170]	82.4	0.85	2021
Ismail et al. [171]	90.73	90.62	2021
Zhang et al. [60]	81.08	0.85	2021
Li, W., and Shen, Z. [55]	83.81	–	2022
Lee et al. [56]	–	76.50	2023
Present study, face Cut-out 2	92	0.93	2024

Table 4.5 Best Performance Comparison of Baseline Deepfake Detection Models on the Celeb Dataset.

In contrasting the methodology of our study with that of Zhang et al. [60], we note that their technique, which entailed random dropout of frame segments, garnered an AUC of 88.83%. This demonstrates commendable efficacy, albeit slightly lower than our approach that achieved an AUC of 93%. Distinctively, the strategy in [171], utilizing the YOLO face detector in tandem with the InceptionResNetV2 CNN, offers a divergent paradigm in DeepFake detection. Their process involves the YOLO detector isolating facial regions from video frames, followed by the InceptionResNetV2 CNN extracting features for subsequent classification by an XGBoost classifier. This approach yielded an AUC of 90.62% and an accuracy of 90.73%.

In our research, the 'Face Cut-out 2' methodology marked a significant advancement, attaining a 93% AUC on the Celeb-DF dataset, thus substantially exceeding the peak accuracy of 83.51% in the FD2Foremer study [55] using their Img+ detail (swin) approach. The latter focuses on analyzing mid-frequency facial geometry details, including individual-specific and dynamic expression-related features.

Another innovative method, introduced in study [56], applies a No-Reference Image Quality Assessment on a patch-by-patch basis, differentiating between facial and non-facial regions, alongside a frequency-decomposition block to extract various frequency components. Although this method achieved an AUC of 76.50% on the Celeb dataset and shows potential, it falls short of the performance attained by our method on the same dataset.

This comparison suggests that our model develops better representations than some of the earlier methods. In conclusion, our study's novel approach of selectively obscuring different facial parts has illuminated the vital role of various facial features in DeepFake detection. The insights derived from this methodology are invaluable for

guiding future research endeavors in enhancing the accuracy and reliability of DeepFake detection models.

4.6.3 Deepfake Detection Approaches that Used Similar Techniques

Deep neural network models are widely utilized for detection purposes, with the choice of specific algorithms varying based on the data type and the insights gained during training. For instance, the FDML model in a study by Liao et al. [172] focuses on detecting fake news by combining fake news detection with news topic classification, employing a unique news graph method and dynamic weighting strategy. In contrast, the "FraudTrip" study by Ding et al. [173] identifies fraudulent taxi trips by analyzing GPS data, diverging from textual content analysis.

In our research, we concentrate on facial image analysis to distinguish DeepFake images, using a technique called "face cut-out" to analyze the importance of different facial regions in DeepFake detection. These diverse applications highlight the adaptability of neural networks in handling various data types and objectives, from multimedia content in fake news to spatial data in fraud detection and facial feature analysis in DeepFake identification.

In our study, we focused on defined face cut-outs to specifically target and analyze the importance of different facial regions in deepfake detection. This targeted approach was inspired by research into prosopagnosia, which highlights the significance of internal facial features such as the eyes and nose for face recognition. Our aim was to investigate whether these insights could be leveraged to enhance deepfake detection accuracy.

We conducted experiments comparing three groups: the baseline group with no cut-outs or augmentation, and two groups using defined face cut-outs targeting specific facial regions. In Cut-Out 1, we covered external facial features (forehead, chin, mouth, and jawline) to make the model focus more on the core features of the face (eyes and nose). In Cut-Out 2, we targeted the core features (eyes and nose) to make the model focus on the external facial features. This group showed higher accuracy compared to the other groups.

To ensure a fair comparison and a better understanding of the specific advantages offered by defined facial region occlusions, we compared the results of our defined face cut-outs with studies that used random cut-outs in the images to improve deepfake

detection and reduce overfitting. The results, presented in Table 4.6, detail the type of occlusions and the accuracy of each study.

A comparative analysis was conducted to demonstrate the effectiveness of our approach in detecting deepfakes compared to other methods using similar techniques. Our model outperformed all other methods that used random cutout , achieving an accuracy of 92%. The second column of the table indicates which parts of the images were covered during training. Notably, approaches employing random cut-outs exhibited the lowest accuracy, while our approach, which utilized a method for selecting specific facial regions to cover, achieved the best performance.

The study by Zhong et al. [151] demonstrated a method involving random rectangular region erasure in images, achieving an accuracy of 76.7%. This approach's relative ineffectiveness might stem from its non-specific targeting of facial features, which are crucial for detecting nuanced manipulations typical in deepfakes. Without focusing on specific facial characteristics, the method possibly overlooks subtle yet critical cues of deepfake alterations.

Han [174] introduced a different technique, using everyday objects like sunglasses, face masks, or hats to occlude parts of the face, achieving a 77.4% accuracy. The relatively modest performance might be attributed to the challenge of differentiating between natural occlusions (like sunglasses in a genuine image) and artificial alterations in deepfakes. This ambiguity could limit the model's ability to accurately identify deepfakes.

DeVries and Taylor [152] explored a method involving random square cut-outs in images, achieving an 81 % accuracy. Again, the randomness of occlusion might have constrained its effectiveness, as it does not specifically address the manipulation of facial features critical in deepfake generation.Similarly, Wang et al. [175] used random erasure methods in their approach, achieving an accuracy of 80.2%. These studies further illustrate the variability in performance when random cut-outs are used.

Another study by Choi and Kim [176], proposed a novel data augmentation technique involving colorful cutouts with curriculum learning. This approach improved model performance by introducing varying levels of noise and difficulty, achieving an accuracy of 78.65%. This finding further confirms that defined cut-outs provide better results than random cut-outs.

The study by Huang et al. [162], which achieved an 87.08% accuracy rate by focusing on occluding three key facial regions—the eyes, nose, and mouth—aligns

closely with our findings. Both their approach and our study highlight the efficacy of placing occlusions in central features of the face. Our study discovered that strategically occluding central facial regions, such as the eyes or nose, substantially improves the model's accuracy in detecting deepfakes. This concurrence between the two studies reinforces the idea that targeting these critical areas of the face is key in developing more efficient deepfake detection models.

Reference	Type of Occlusion	Accuracy (%)
[151]	Randomly selecting and erasing a rectangular region in the image	76.7
[152]	Random square cut-out in the image	81.0
[174]	Random objects (such as sunglasses, face mask, or hat) to occlude different parts of the face	77.4
[175]	Random cut-out in the image	80.02
[176]	Random colorful square cutouts	78.65
[162]	Feature-integration approach for occlusion of three regions: eyes, nose, and mouth	87.08
Our approach	Defined cut-out: left eye, right eye, both eyes, or nose	**92**

Table 4.6 Comparative Results of State-of-the-Art Deepfake Detection Techniques Using Different Occlusion Methods

Overall, these studies collectively highlight the importance of targeted and strategic approaches in deepfake detection. The varying levels of success underline the need for methods that focus on critical facial features and subtle cues to improve the accuracy and reliability of deepfake detection technologies. Inspiration for our method was drawn from the medical field, particularly the study of prosopagnosia, a condition that impairs the ability to recognize faces. The investigation of this condition provided insight into the facial regions most critical for face recognition, which are targeted in medical training and rehabilitation programs. Therefore, the selection of face cut-outs in each group was based on insights from prosopagnosia research.

In Cut-Out 1, we occluded external regions to keep the eyes and nose regions visible, typically targeting these core regions in training programs to improve the ability of individuals with prosopagnosia to identify faces. This group showed lower accuracy. In Cut-Out 2, we focused on covering the eyes and nose regions to examine the effect on model performance. This group showed higher accuracy, even after covering the core

regions of the face. This finding suggests that the AI model operates differently from individuals with prosopagnosia.

Leveraging this knowledge, facial regions were selected for occlusion with the aim of improving the effectiveness of the deepfake detection model. Overall, the findings suggest that the proposed approach has the potential to enhance the accuracy and reliability of deepfake detection methods.

4.7 Chapter Summary

This chapter has presented the results achieved by training the model used in this study with the FaceForensics++ and Celeb-DF datasets, and compared these results with the findings of other research. The models were trained on both datasets in three different settings: (1) without image augmentation; (2) with face Cut-out 1 augmentation; and (3) with face Cut-out 2 augmentation. The results of each phase of the experiment were analyzed, and it was found that the model trained with face Cut-out 2 augmentation outperformed the other two conditions in both phases 1 and 2. Moreover, during Phase 3, an independent assessment of the Cutout Technique was conducted for each individual facial feature, revealing that the nose played the least significant role in deepfake detection.

To ensure a fair comparison and gain a better understanding of the specific advantages offered by defined facial region occlusions, we compared the results of our defined face cut-outs with studies that used random cut-outs in the images to improve deepfake detection and reduce overfitting. This allowed us to contextualize the unique contributions of our approach within the broader field of deepfake detection research.

In addition, the results of the approach used in this study were compared with those of state-of-the-art methods that used the FF++ datasets and Celeb-DF datasets, as well as similar techniques with different datasets, in order to ensure a comprehensive comparison. In the next chapter, a brief summary of the key findings of the present study is provided, its limitations are identified, and recommendations are made for future research.

Chapter 5

Conclusion

5.1 Introduction

In this final chapter, we bring this dissertation to a close by revisiting the research questions posed in Chapter 1, summarising the key findings, and outlining potential avenues for future research. Our emphasis will be on the main findings and overarching insights, while the specific details can be found in the conclusion sections of the respective chapters.

5.2 Thesis Summary

This thesis investigates the intersection of deepfake detection and prosopagnosia, shedding light on both fields. The comparison between our deepfake detection models and insights from prosopagnosia studies reveals intriguing parallels and key differences that highlight the complexities of facial recognition in both artificial intelligence and human cognition. Unlike deep neural models, which allow for precise control over experimental variables, the study of prosopagnosia presents challenges due to the unique and varied nature of brain damage in affected individuals. This makes it difficult to maintain consistent conditions when researching the condition, unlike the controlled settings of AI models.

The research demonstrates that AI models, specifically EfficientNet-B7 and XceptionNet, performed best with the Face Cut-out 2 technique, which occludes core facial features like the eyes and nose, compared to Face Cut-out 1, which focuses on external

features, and the baseline model with no occlusions. These findings suggest that targeting non-core facial features can be more effective for deepfake detection. Additionally, the models showed stronger performance on the Celeb-DF dataset compared to the FF++ dataset, consistent with findings by Yang et al. [61], Haliassos et al. [170], and Lee et al. [56]. However, studies such as Bonettini et al. [177] reported better results with the FF++ dataset, indicating that detection performance can vary based on the techniques and models used.

In Phase 3, which focused on occluding individual facial features, the results showed that obscuring the nose and mouth improved detection performance, while covering the eyes led to poorer outcomes. This highlights the varying importance of different facial features in deepfake detection.

Overall, the Face Cut-out technique enhances deepfake detection by emphasizing critical facial features during training and helps reduce overfitting. This method shows significant potential for broader applications in improving the robustness and accuracy of deepfake detection systems.

5.3 Comparative Analysis of Deepfake Detection and Prosopagnosia: Identifying Similarities and Differences

After studying the condition of prosopagnosia, we discovered that most training programs often shift the focus of the patients from external parts of the face, such as the chin and forehead, to more central features like the eyes and nose. This shift led to an improvement in patients' ability to recognize faces. Inspired by this, we applied a similar approach in our dataset to train a deep neural network model to identify deepfakes.

We divided the training into three groups. In the first group, we covered the internal features of the faces, like the eyes and nose, to encourage the model to focus on the external features. For the second group, we did the opposite, covering the external features and leaving only the internal parts visible. The third group, our baseline, had no alterations to the facial features.

After training the model with these three distinct groups, we tested it on unseen data to assess its ability to differentiate between real and fake faces. Interestingly, the

results showed that the model performed best with the first group, where the internal facial features were covered. This suggests that external facial features might provide more significant clues for detecting deepfakes.

5.3.1 Details of Applying Insights from Prosopagnosia to Deepfake Technology

Our findings present an interesting dynamic in how different facial features influence the performance of the deepfake detection model. Let's break down the two scenarios:

- Covering Multiple Facial Features: when multiple facial regions were covered within one group, the group where internal features (such as the eyes and nose) were obscured demonstrated higher accuracy compared to the group with external features (like the hairline, ears, jawline) covered. This indicates that, generally, the model might depend more on external features for deepfake detection when multiple areas are obscured.
- Covering Individual Facial Features: However, when testing each facial feature individually, we observed a different pattern. The model's performance was most significantly impacted when the eyes were covered, indicating a decrease in accuracy. Conversely, covering the nose resulted in the best performance.

5.3.2 Similarities and Differences Between Deepfake Detection and Prosopagnosia

The results from our deepfake detection model and the insights from prosopagnosia studies show some intriguing parallels, but they also reveal differences that highlight the complexities of facial recognition in both artificial and human systems. Here's how they align and differ:

Alignment in the Importance of Specific Features

The approach to facial recognition in both deepfake detection models and individuals with prosopagnosia is grounded in a similar fundamental principle. In each case, the face is a critical element in the recognition process. These models or individuals perceive faces as broadly alike, and they actively seek specific features or clues to

differentiate one face from another. Fundamentally, they both operate on the principle of identifying distinct facial characteristics, a concept that is central to research in these fields.

Our deepfakes model results, paralleled with findings from prosopagnosia research, highlight the significance of particular facial features, such as the eyes, in recognition tasks. The notable decline in our model's performance when the eyes were obscured reflects similar recognition challenges experienced by individuals with prosopagnosia. They often face difficulties in recognizing faces due to impaired processing of key facial features, including the eyes.

Effective differentiation between faces, for both deepfake detection models and individuals with prosopagnosia, requires specialized training in facial recognition techniques.

Variations in Feature Reliance and Adaptation Strategies

In prosopagnosia, training that focuses on recognizing internal facial features, like the eyes and nose, improves individuals' face recognition abilities. Typically, before this training, these individuals depend more on external features or non-facial cues, as they find it challenging to process internal facial details. Conversely, our model showed a different response. When trained to focus on internal features by hiding external ones, its accuracy decreased. However, the model's accuracy improved when the internal features were covered, leaving external features visible. This indicates that, unlike individuals with prosopagnosia, our model is more effective at recognizing deepfakes using external facial features.

5.3.3 Exploring the Possible Causes of the Differences

Our deep neural model has very specific settings that we can adjust. This means we can control our experiments so that the results are only affected by the things we want to study. On the other hand, with prosopagnosia, which is a condition affecting face recognition, it's hard to control what's happening in the brain. People with prosopagnosia can have different types of brain damage, making each case unique. So, unlike our model, it's much harder to have consistent conditions when studying prosopagnosia.

Human Facial Recognition: In humans, the process of recognizing faces is complex and largely driven by neurological factors. This is particularly noticeable in individuals with conditions like prosopagnosia, who often need to adjust their recognition strategies. They might shift their focus to different facial features, adapting based on their cognitive abilities and limitations.

AI Facial Recognition: For AI models, the complexity in facial recognition stems from the data and algorithms used. Our model demonstrates adaptability similar to humans by changing its focus on various facial features based on its training. However, the basis of this adaptability in AI is different from that in humans. It's grounded in the way the model processes data and makes decisions algorithmically, not in neurological functioning.

Inconclusion, The decision to apply strategies from prosopagnosia research to improve deepfake detection models is influenced by the contrasting research histories of these two fields. Prosopagnosia has been extensively studied since the 1940s, enriching neurology and cognitive science with a wealth of knowledge and a comprehensive literature base. This extensive research has provided a deep, foundational understanding of facial recognition challenges in prosopagnosia. On the other hand, deepfake technology, emerging prominently around 2017, is still in its infancy, with many areas yet to be fully explored and understood.

Given the rich insights and well-established strategies in prosopagnosia research, we believe these can be effectively adapted to enhance AI, particularly in the realm of deepfake detection. The depth of understanding in prosopagnosia offers valuable lessons and techniques that could address some of the nascent challenges in the rapidly evolving field of deepfake technology. Therefore, we advocate for integrating medical field methodologies, especially those inspired by neurological studies of the human brain, into future AI research to develop more robust and effective solutions for deepfake detection.

5.4 Limitations

5.4.1 Dataset

One prominent limitation is the need for high-quality training data, encompassing genuine and manipulated samples, to build accurate deep fake detection models.

However, obtaining diverse and comprehensive training datasets can be daunting, particularly when dealing with emerging types of deep fakes or specific contextual variations.

Another challenge we faced was the prevalence of imbalanced datasets. In many cases, the number of fake faces outweighs the number of real faces in the available datasets. This imbalance necessitates additional steps to balance the dataset effectively, ensuring a fair representation of actual and manipulated samples.

5.4.2 Dataset Preparation Complexity

Preparing the dataset for training purposes requires a considerable amount of time due to the multiple involved steps. These steps are time-consuming, and some of them even require manual intervention. For instance, after the initial face detection, it becomes necessary to carefully examine the images and extract the face from the background. This is because the face detection algorithm may occasionally extract certain background elements that resemble a face but are, in fact, not.

Moreover, the training process itself is a time-intensive endeavour, particularly when dealing with a large dataset. Additionally, it becomes essential to repeat the process multiple times to ensure an accurate average accuracy is achieved at certain stages. This repetitive nature further extends the overall duration of the training process.

5.4.3 lack of Resources and Information

In my research on deepfake detection, I encountered several challenges. One of the most significant challenges was the lack of clarity in existing studies. This is because the field of deepfake detection is still relatively new, and there are not many studies available.

Additionally, many studies do not clearly explain their methods. This is especially true for the steps involved in preparing the dataset for training. As a result, it is difficult to understand how the methods work and to replicate the results.

Furthermore, many studies do not clearly explain the rationale behind key decisions, such as the selection of computational models, the use of specific techniques, and the

choice of datasets. This makes it difficult to understand why the results were obtained and to apply the methods to other datasets.

5.4.4 Limitations in the Algorithm, Methodology, and Experiments

One key limitation is the reliance of deepfake detection algorithms on the quality and diversity of training data. Like many others, our study utilized deepfake datasets with limited access to high-quality data. This restriction hampers the generalization ability of our models, particularly in detecting sophisticated deepfake techniques.

Furthermore, our experiment is tailored to detect deepfakes in static images, which may restrict the extrapolation of its effectiveness to other media types, such as videos or live streams.

Additionally, our methodology involves selecting cut-out regions in images inspired by training techniques used for prosopagnosia patients. Although we drew upon extensive medical literature, including case studies of individuals with this disorder, the rarity of prosopagnosia and the variability in symptom severity limit the generalizability of findings from individual case studies, as evidenced by the diverse range of abilities in identifying faces observed among prosopagnosia patients.

Moreover, while we tested our models with defined face cut-outs and no cut-outs, we did not test against random cut-outs of the same size/statistics. This omission limits the comprehensiveness of our results, as a fair comparison would involve random cut-outs to determine whether the performance improvements are truly due to the specific defined cut-out strategy or simply any cut-out approach. Future studies should incorporate random cut-outs to provide a more robust comparison and a clearer understanding of the effectiveness of the defined cut-out technique. However, to over come this omission in the current study, we compared our results with those of other studies that utilized random cut-outs.

Future research may also directly involve individuals with prosopagnosia in experiments, providing more robust and insightful data. By including individuals with similar levels of the disorder, we can gain a clearer understanding of its impact and evaluate their performance using our dataset.

Mitigating these limitations is essential for advancing the field of deepfake detection and minimizing the risks associated with synthetic media proliferation. Future endeav-

ors should prioritize collaborative efforts to overcome these challenges and ensure the ongoing effectiveness of deepfake detection technologies.

5.5 Future Work Recommendations

While this thesis introduces a novel face cutout technique designed to enhance deepfake detection methods, it also paves the way for several promising directions to extend the research presented herein. In particular, these directions can be categorized into two main areas:

5.5.1 Technological Advancements

Moving forward, our future endeavors will revolve around expanding the application of this augmentation policy to various face manipulation and forgery datasets. We also aim to explore its efficacy with different architectures of deep neural models, thereby encompassing a wider scope of experiments.

Furthermore, our primary focus in future research will be on enhancing the generalization capability of deepfake detection models. We intend to dedicate efforts towards improving the models' ability to detect deep fakes in various scenarios and contexts.

Finally, our future work involves conducting more detailed analyses aimed at identifying facial regions that contribute minimally to information processing. This will involve a methodical approach, where we will selectively obscure areas like the nose and mouth, which preliminary findings suggest are less informative. We will then assess whether this selective obscuration improves accuracy compared to a previous experimental groups . This comparative analysis will help us understand the impact of different facial regions on the overall effectiveness of our model.Through more experiments, we aim to gain insights into the effectiveness of focusing on these specific areas when it comes to deepfake detection, allowing us to refine our methodologies and strategies. We believes our research can useful insights for the research community by Improving the next generation of deepfakes by focusing on the artifacts that exist in specific facial regions.

In conclusion, this study underscores the significance of integrating multidisciplinary knowledge, including medical expertise, to advance the detection capabilities in the realm of deepfake technology. It is widely acknowledged among researchers that the

dynamic between facial manipulation techniques and their detection is in a state of perpetual evolution, where advancements in one domain catalyze progress in the other. The incorporation of diverse knowledge bases, particularly from the medical field, offers a promising avenue to disrupt this ongoing cycle. By amalgamating medical insights with conventional detection methodologies, there is a substantial potential to develop more sophisticated and resilient deepfake detection strategies. This interdisciplinary approach could be pivotal in staying ahead in the increasingly complex landscape of digital content verification.

5.5.2 Medical Applications

The medical field represents a promising area for future research applications of the cut-out automation technique developed in this study. This technique has potential utility in assisting doctors with the creation of personalized perceptual training exercises using images with occlusions covering certain regions of the face. These exercises, which concentrate on distinct facial features such as the eyes, nose, and mouth, could prove valuable for the rehabilitation or training of patients with prosopagnosia (face blindness).

By employing these techniques, doctors are able to design customized training datasets tailored to the specific needs of each patient. These datasets, consisting of images with targeted face cuts, can be distributed and accessed online, thus facilitating remote training. This approach offers significant advantages in terms of global patient care. It is particularly beneficial for individuals with less common conditions requiring prolonged treatment, such as prosopagnosia, where specialized and consistent training is crucial for improvement.

Facial cut-out methods offer a granular approach to analyzing facial features, enabling targeted exploration of specific regions. Face Cutout 1 isolates core facial features by occluding external elements such as the forehead, chin, mouth, or jawline. Conversely, Face Cutout 2 focuses on external features by masking core components like the eyes or nose. This flexibility empowers researchers to formulate and test hypotheses, opening new avenues for investigating the contributions of different facial features to perception and recognition. By tailoring training exercises to specific facial regions, clinicians can provide personalized rehabilitation for patients with prosopagnosia, focusing on improving recognition of impaired features such as the eyes or mouth.

Additionally, our method can be utilized to conceal specific parts of a patient's face, thereby protecting their identity. This feature is particularly useful in medical training scenarios where students need to study certain facial features without compromising patient privacy. By selectively revealing only the relevant parts of the face, our technique ensures that patient confidentiality is maintained while providing valuable educational material for medical students.

Finally, Our method also increases the number of medical images available for research. This leads to larger, more diverse sample sizes in studies, thereby broadening the knowledge base in specialized fields. Such an enhancement not only enriches scientific literature but also provides substantial benefits to the medical community at large.

In terms of future prospects, the applications of our techniques have the potential to advance medical research and education. By developing and refining these approaches, we aim to contribute to improvements in medical science and patient care. Our work highlights the importance of continued innovation and collaboration in the field of healthcare.

Bibliography

[1] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "Deepfakes and beyond: A survey of face manipulation and fake detection," *Information Fusion*, vol. 64, pp. 131–148, 2020.

[2] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*, pp. 6105–6114, PMLR, 2019.

[3] J. E. Arco, J. Ramírez, J. M. Górriz, M. Ruz, A. D. N. Initiative, *et al.*, "Data fusion based on searchlight analysis for the prediction of alzheimer's disease," *Expert Systems with Applications*, vol. 185, p. 115549, 2021.

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

[5] B. Chesney and D. Citron, "Deep fakes: A looming challenge for privacy, democracy, and national security," *Calif. L. Rev.*, vol. 107, p. 1753, 2019.

[6] V. A. Jones, *Artificial intelligence enabled deepfake technology: The emergence of a new threat*. PhD thesis, Utica College, 2020.

[7] T. Bylemans, L. Vrancken, and K. Verfaillie, "Developmental prosopagnosia and elastic versus static face recognition in an incidental learning task," *Frontiers in Psychology*, vol. 11, p. 2098, 2020.

[8] P. Korshunov and S. Marcel, "Deepfakes: a new threat to face recognition? assessment and detection," *arXiv preprint arXiv:1812.08685*, 2018.

[9] B. Sorger, R. Goebel, C. Schiltz, and B. Rossion, "Understanding the functional neuroanatomy of acquired prosopagnosia," *Neuroimage*, vol. 35, no. 2, pp. 836–852, 2007.

[10] M. Masood, M. Nawaz, K. M. Malik, A. Javed, A. Irtaza, and H. Malik, "Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward," *Applied Intelligence*, vol. 53, no. 4, pp. 3974–4026, 2023.

[11] P. N. Vasist and S. Krishnan, "Deepfakes: an integrative review of the literature and an agenda for future research," *Communications of the Association for Information Systems*, vol. 51, no. 1, p. 14, 2022.

[12] I. R. Khan, S. Aisha, D. Kumar, and T. Mufti, "A systematic review on deepfake technology," *Proceedings of Data Analytics and Management: ICDAM 2022*, pp. 669–685, 2023.

[13] S. Subramanya and B. K. Yi, "Digital rights management," *IEEE potentials*, vol. 25, no. 2, pp. 31–34, 2006.

[14] D. Dagar and D. K. Vishwakarma, "A literature review and perspectives in deepfakes: generation, detection, and applications," *International Journal of Multimedia Information Retrieval*, vol. 11, no. 3, pp. 219–289, 2022.

[15] T. D. Akinosho, L. O. Oyedele, M. Bilal, A. O. Ajayi, M. D. Delgado, O. O. Akinade, and A. A. Ahmed, "Deep learning in the construction industry: A review of present status and future innovations," *Journal of Building Engineering*, vol. 32, p. 101827, 2020.

[16] J. Bijhold, A. Ruifrok, M. Jessen, Z. Geradts, S. Ehrhardt, and I. Alberink, "Forensic audio and visual evidence 2004-2007: A review," in *15th INTERPOL forensic science symposium*, 2007.

[17] B. R. Ulbricht, C. Moxley, M. D. Austin, and M. D. Norburg, "Digital eye-witnesses: Using new technologies to authenticate evidence in human rights litigation," *Stan. L. Rev.*, vol. 74, p. 851, 2022.

[18] S. Karnouskos, "Artificial intelligence in digital media: The era of deepfakes," *IEEE Transactions on Technology and Society*, vol. 1, no. 3, pp. 138–147, 2020.

[19] MAGIX Software GmbH, "Sound forge." https://www.magix.com/gb/music/sound-forge/, 2021. Accessed January 11, 2021.

[20] "Faceapp." https://www.faceapp.com/, 2021. Accessed September 17, 2021.

[21] Reface Team, "Reface app." https://reface.app/, 2021. Accessed September 11, 2021.

[22] Audacity Team, "Audacity." https://www.audacityteam.org/, 2021. Accessed September 09, 2021.

[23] T. T. Nguyen, Q. V. H. Nguyen, D. T. Nguyen, D. T. Nguyen, T. Huynh-The, S. Nahavandi, T. T. Nguyen, Q.-V. Pham, and C. M. Nguyen, "Deep learning for deepfakes creation and detection: A survey," *Computer Vision and Image Understanding*, vol. 223, p. 103525, 2022.

[24] S. Alanazi and S. Asif, "Understanding deepfakes: A comprehensive analysis of creation, generation, and detection," *Artificial Intelligence and Social Computing*, vol. 72, no. 72, 2023.

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

[26] C. Bregler, M. Covell, and M. Slaney, "Video rewrite: Driving visual speech with audio," in *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, pp. 353–360, 1997.

[27] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 23, no. 6, pp. 681–685, 2001.

[28] 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*, pp. 2387–2395, 2016.

[29] R. Winter and A. Salter, "Deepfakes: uncovering hardcore open source on github," *Porn Studies*, vol. 7, no. 4, pp. 382–397, 2020.

[30] E. Meskys, J. Kalpokiene, P. Jurcys, and A. Liaudanskas, "Regulating deep fakes: legal and ethical considerations," *Journal of Intellectual Property Law & Practice*, vol. 15, no. 1, pp. 24–31, 2020.

[31] P. Stokes, "Ghosts in the machine: Do the dead live on in facebook?," *Philosophy & Technology*, vol. 25, no. 3, pp. 363–379, 2012.

[32] M. A. Radetzky, B. Butr-Indr, *et al.*, *Deep learning technology and its impact on a cinematographic work*. PhD thesis, Thammasat University, 2023.

[33] Y. Mirsky and W. Lee, "The creation and detection of deepfakes: A survey," *ACM Computing Surveys (CSUR)*, vol. 54, no. 1, pp. 1–41, 2021.

[34] V. Jaiman and V. Urovi, "A consent model for blockchain-based health data sharing platforms," *IEEE Access*, vol. 8, pp. 143734–143745, 2020.

[35] C. Vaccari and A. Chadwick, "Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news.," *Social Media+ Society*, vol. 6, no. 1, 2020.

[36] T. Paterson and L. Hanley, "Political warfare in the digital age: cyber subversion, information operations and 'deep fakes'," *Australian Journal of International Affairs*, vol. 74, no. 4, pp. 439–454, 2020.

[37] A. De Ruiter, "The distinct wrong of deepfakes," *Philosophy & Technology*, vol. 34, no. 4, pp. 1311–1332, 2021.

[38] D. K. Sharma, L. Gaur, and D. Okunbor, "Image compression and feature extraction with neural network," in *Allied Academies International Conference. Academy of Management Information and Decision Sciences. Proceedings*, vol. 11, p. 33, Citeseer, 2007.

[39] G. Singh, B. Kumar, L. Gaur, and A. Tyagi, "Comparison between multinomial and bernoulli naïve bayes for text classification," in *2019 International Conference on Automation, Computational and Technology Management (ICACTM)*, pp. 593–596, IEEE, 2019.

[40] L. Gaur, G. K. Arora, and N. Z. Jhanjhi, "Deep learning techniques for creation of deepfakes," in *DeepFakes*, pp. 23–34, CRC Press, 2022.

[41] A. K. Singh, "Deep learning for deepfakes creation and detection," 2024.

[42] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410, 2019.

[43] K. Lin, W. Han, Z. Gu, and S. Li, "A survey of deepfakes generation and detection," in *2021 IEEE Sixth International Conference on Data Science in Cyberspace (DSC)*, pp. 474–478, IEEE, 2021.

[44] Intel, "faceswap-gan," Nov 2022.

[45] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *British Machine Vision Conference*, 2015.

[46] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.

[47] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman, "Synthesizing obama: learning lip sync from audio," *ACM Transactions on Graphics (ToG)*, vol. 36, no. 4, pp. 1–13, 2017.

[48] P. Korshunov and S. Marcel, "Speaker inconsistency detection in tampered video," in *2018 26th European signal processing conference (EUSIPCO)*, pp. 2375–2379, IEEE, 2018.

[49] A. A. Chien, "Open collaboration in an age of distrust," 2018.

[50] E. Kakaletsis and N. Nikolaidis, "A technique for fake 3d (2d-to-3d converted) video recognition," in *2015 23rd European Signal Processing Conference (EUSIPCO)*, pp. 106–109, IEEE, 2015.

[51] 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*, pp. 81–87, 2018.

[52] C.-C. Hsu, C.-Y. Lee, and Y.-X. Zhuang, "Learning to detect fake face images in the wild," in *2018 international symposium on computer, consumer and control (IS3C)*, pp. 388–391, IEEE, 2018.

[53] M. T. Jafar, M. Ababneh, M. Al-Zoube, and A. Elhassan, "Forensics and analysis of deepfake videos," in *2020 11th international conference on information and communication systems (ICICS)*, pp. 053–058, IEEE, 2020.

[54] U. A. Ciftci, I. Demir, and L. Yin, "Fakecatcher: Detection of synthetic portrait videos using biological signals," *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[55] W. Li, Z. Shen, *et al.*, "Fd 2 foremer: Thinking face forgery detection in mid-frequency geometry details," *Security and Communication Networks*, vol. 2022, 2022.

[56] E.-G. Lee, I. Lee, and S.-B. Yoo, "Cluecatcher: Catching domain-wise independent clues for deepfake detection," *Mathematics*, vol. 11, no. 18, p. 3952, 2023.

[57] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan, "Recurrent convolutional strategies for face manipulation detection in videos," *Interfaces (GUI)*, vol. 3, no. 1, pp. 80–87, 2019.

[58] I. Amerini and R. Caldelli, "Exploiting prediction error inconsistencies through lstm-based classifiers to detect deepfake videos," in *Proceedings of the 2020 ACM workshop on information hiding and multimedia security*, pp. 97–102, 2020.

[59] T. Jung, S. Kim, and K. Kim, "Deepvision: Deepfakes detection using human eye blinking pattern," *IEEE Access*, vol. 8, pp. 83144–83154, 2020.

[60] D. Zhang, C. Li, F. Lin, D. Zeng, and S. Ge, "Detecting deepfake videos with temporal dropout 3dcnn.," in *IJCAI*, pp. 1288–1294, 2021.

[61] X. Yang, Y. Li, and S. Lyu, "Exposing deep fakes using inconsistent head poses," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8261–8265, IEEE, 2019.

[62] D. Menotti, G. Chiachia, A. Pinto, W. R. Schwartz, H. Pedrini, A. X. Falcao, and A. Rocha, "Deep representations for iris, face, and fingerprint spoofing detection," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 4, pp. 864–879, 2015.

[63] H. Steiner, S. Sporrer, A. Kolb, and N. Jung, "Design of an active multispectral swir camera system for skin detection and face verification," *Journal of Sensors*, vol. 2016, 2016.

[64] X. Li, J. Komulainen, G. Zhao, P.-C. Yuen, and M. Pietikäinen, "Generalized face anti-spoofing by detecting pulse from face videos," in *2016 23rd International Conference on Pattern Recognition (ICPR)*, pp. 4244–4249, IEEE, 2016.

[65] S. Rana, S. Gaj, A. Sur, and P. K. Bora, "Detection of fake 3d video using cnn," in *2016 IEEE 18th International Workshop on Multimedia Signal Processing (MMSP)*, pp. 1–5, IEEE, 2016.

[66] G. R. Kumar, R. K. Kumar, and G. Sanyal, "Facial emotion analysis using deep convolution neural network," in *2017 International Conference on Signal Processing and Communication (ICSPC)*, pp. 369–374, IEEE, 2017.

[67] K. Mhou, D. van der Haar, and W. S. Leung, "Face spoof detection using light reflection in moderate to low lighting," in *2017 2nd Asia-Pacific Conference on Intelligent Robot Systems (ACIRS)*, pp. 47–52, IEEE, 2017.

[68] Y. Li, M.-C. Chang, and S. Lyu, "In ictu oculi: Exposing ai created fake videos by detecting eye blinking," in *2018 IEEE international workshop on information forensics and security (WIFS)*, pp. 1–7, IEEE, 2018.

[69] D. Güera, S. Baireddy, P. Bestagini, S. Tubaro, and E. J. Delp, "We need no pixels: Video manipulation detection using stream descriptors," *arXiv preprint arXiv:1906.08743*, 2019.

[70] L. Guarnera, O. Giudice, and S. Battiato, "Deepfake detection by analyzing convolutional traces," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 666–667, 2020.

[71] D. M. Montserrat, H. Hao, S. K. Yarlagadda, S. Baireddy, R. Shao, J. Horváth, E. Bartusiak, J. Yang, D. Guera, F. Zhu, *et al.*, "Deepfakes detection with automatic face weighting," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 668–669, 2020.

[72] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, pp. 1359–1367, 2020.

[73] M. Lee, Y. K. Lee, M.-T. Lim, and T.-K. Kang, "Emotion recognition using convolutional neural network with selected statistical photoplethysmogram features," *Applied Sciences*, vol. 10, no. 10, p. 3501, 2020.

[74] B. Rossion, C. Schiltz, L. Robaye, D. Pirenne, and M. Crommelinck, "How does the brain discriminate familiar and unfamiliar faces?: a pet study of face categorical perception," *Journal of cognitive neuroscience*, vol. 13, no. 7, pp. 1019–1034, 2001.

[75] V. Bruce and A. Young, "Understanding face recognition," *British journal of psychology*, vol. 77, no. 3, pp. 305–327, 1986.

[76] C.-C. Chen, K.-L. C. Kao, and C. W. Tyler, "Face configuration processing in the human brain: The role of symmetry," *Cerebral Cortex*, vol. 17, no. 6, pp. 1423–1432, 2007.

[77] C. A. Nelson, "The development and neural bases of face recognition," *Infant and Child Development: An International Journal of Research and Practice*, vol. 10, no. 1-2, pp. 3–18, 2001.

[78] O. Pascalis, X. de Martin de Viviés, G. Anzures, P. C. Quinn, A. M. Slater, J. W. Tanaka, and K. Lee, "Development of face processing," *Wiley Interdisciplinary Reviews: Cognitive Science*, vol. 2, no. 6, pp. 666–675, 2011.

[79] N. K. Logothetis, J. Pauls, and T. Poggio, "Shape representation in the inferior temporal cortex of monkeys," *Current biology*, vol. 5, no. 5, pp. 552–563, 1995.

[80] T. Kress and I. Daum, "Developmental prosopagnosia: A review," *Behavioural neurology*, vol. 14, no. 3-4, pp. 109–121, 2003.

[81] J. Geskin and M. Behrmann, "Congenital prosopagnosia without object agnosia? a literature review," *The Face Specificity of Lifelong Prosopagnosia*, pp. 4–54, 2020.

[82] J. M. Henderson, C. C. Williams, and R. J. Falk, "Eye movements are functional during face learning," *Memory & cognition*, vol. 33, no. 1, pp. 98–106, 2005.

[83] M. Krickl, U. Poser, and H. J. Markowitsch, "Interactions between damaged brain hemisphere and mode of presentation on the recognition of faces and figures," *Neuropsychologia*, vol. 25, no. 5, pp. 795–805, 1987.

[84] E. Mayer and B. Rossion, *Prosopagnosia*, p. 315–334. Cambridge University Press, 2007.

[85] N. Mestry, N. Donnelly, T. Menneer, and R. A. McCarthy, "Discriminating thatcherised from typical faces in a case of prosopagnosia," *Neuropsychologia*, vol. 50, no. 14, pp. 3410–3418, 2012.

[86] M. Pizzamiglio, M. De Luca, A. Di Vita, L. Palermo, A. Tanzilli, C. Dacquino, and L. Piccardi, "Congenital prosopagnosia in a child: Neuropsychological assessment, eye movement recordings and training," *Neuropsychological rehabilitation*, vol. 27, no. 3, pp. 369–408, 2017.

[87] J. Davidoff and T. Landis, "Recognition of unfamiliar faces in prosopagnosia," *Neuropsychologia*, vol. 28, no. 11, pp. 1143–1161, 1990.

[88] S. Berent, "Functional asymmetry of the human brain in the recognition of faces," *Neuropsychologia*, vol. 15, no. 6, pp. 829–831, 1977.

[89] A. Stone and T. Valentine, "Perspectives on prosopagnosia and models of face recognition," *Cortex*, vol. 39, no. 1, pp. 31–40, 2003.

[90] A. M. Burton, A. W. Young, V. Bruce, R. A. Johnston, and A. W. Ellis, "Understanding covert recognition," *Cognition*, vol. 39, no. 2, pp. 129–166, 1991.

[91] K. Falahkheirkhah, S. Tiwari, K. Yeh, S. Gupta, L. Herrera-Hernandez, M. R. McCarthy, R. E. Jimenez, J. C. Cheville, and R. Bhargava, "Deepfake histologic images for enhancing digital pathology," *Laboratory Investigation*, vol. 103, no. 1, p. 100006, 2023.

[92] K. W. Greve and R. M. Bauer, "Implicit learning of new faces in prosopagnosia: An application of the mere-exposure paradigm," *Neuropsychologia*, vol. 28, no. 10, pp. 1035–1041, 1990.

[93] J. N. Van Der Geest, C. Kemner, M. N. Verbaten, and H. Van Engeland, "Gaze behavior of children with pervasive developmental disorder toward human faces: a fixation time study," *Journal of child psychology and psychiatry*, vol. 43, no. 5, pp. 669–678, 2002.

[94] R. Schwarzer, B. Schüz, J. P. Ziegelmann, S. Lippke, A. Luszczynska, and U. Scholz, "Adoption and maintenance of four health behaviors: Theory-guided longitudinal studies on dental flossing, seat belt use, dietary behavior, and physical activity," *Annals of behavioral medicine*, vol. 33, no. 2, pp. 156–166, 2007.

[95] R. J. Bennetts, N. Butcher, K. Lander, R. Udale, and S. Bate, "Movement cues aid face recognition in developmental prosopagnosia.," *Neuropsychology*, vol. 29, no. 6, p. 855, 2015.

[96] A. K. Bobak, B. A. Parris, N. J. Gregory, R. J. Bennetts, and S. Bate, "Eye-movement strategies in developmental prosopagnosia and "super" face recognition," *Quarterly journal of experimental psychology*, vol. 70, no. 2, pp. 201–217, 2017.

[97] S. Bate, R. Bennetts, J. A. Mole, J. A. Ainge, N. J. Gregory, A. K. Bobak, and A. Bussunt, "Rehabilitation of face-processing skills in an adolescent with prosopagnosia: Evaluation of an online perceptual training programme," *Neuropsychological rehabilitation*, vol. 25, no. 5, pp. 733–762, 2015.

[98] S. Bate and R. J. Bennetts, "The rehabilitation of face recognition impairments: a critical review and future directions," *Frontiers in human neuroscience*, vol. 8, p. 491, 2014.

[99] N. Burra, D. Kerzel, and M. Ramon, "Gaze-cueing requires intact face processing–insights from acquired prosopagnosia," *Brain and cognition*, vol. 113, pp. 125–132, 2017.

[100] A. L. Diaz, "Do i know you? a case study of prosopagnosia (face blindness)," *The Journal of School Nursing*, vol. 24, no. 5, pp. 284–289, 2008.

[101] R. Caldara, P. Schyns, E. Mayer, M. L. Smith, F. Gosselin, and B. Rossion, "Does prosopagnosia take the eyes out of face representations? evidence for a defect in representing diagnostic facial information following brain damage," *Journal of cognitive neuroscience*, vol. 17, no. 10, pp. 1652–1666, 2005.

[102] D. R. Fine, "A life with prosopagnosia," *Cognitive neuropsychology*, vol. 29, no. 5-6, pp. 354–359, 2012.

[103] A. Adams, P. J. Hills, R. J. Bennetts, and S. Bate, "Coping strategies for developmental prosopagnosia," *Neuropsychological Rehabilitation*, vol. 30, no. 10, pp. 1996–2015, 2020.

[104] A. R. Damasio, "Prosopagnosia," *Trends in Neurosciences*, vol. 8, pp. 132–135, 1985.

[105] M. Coltheart, R. Brunsdon, L. Nickels, *et al.*, "Cognitive rehabilitation and its relationship to cognitive-neuropsychological rehabilitation," *Effectiveness of rehabilitation for cognitive deficits*, pp. 11–20, 2005.

[106] L. Schmalzl, R. Palermo, M. Green, R. Brunsdon, and M. Coltheart, "Training of familiar face recognition and visual scan paths for faces in a child with congenital prosopagnosia," *Cognitive Neuropsychology*, vol. 25, no. 5, pp. 704–729, 2008.

[107] R. Brunsdon, M. Coltheart, L. Nickels, and P. Joy, "Developmental prosopagnosia: A case analysis and treatment study," *Cognitive Neuropsychology*, vol. 23, no. 6, pp. 822–840, 2006.

[108] E. Beyn and G. Knyazeva, "The problem of prosopagnosia," *Journal of neurology, neurosurgery, and psychiatry*, vol. 25, no. 2, p. 154, 1962.

[109] B. A. Wilson, *Rehabilitation of memory.* Guilford Press, 1987.

[110] H. D. Ellis and A. W. Young, "Training in face-processing skills for a child with acquired prosopagnosia," *Developmental Neuropsychology*, vol. 4, no. 4, pp. 283–294, 1988.

[111] R. Francis, M. J. Riddoch, and G. W. Humphreys, "'who's that girl?'prosopagnosia, person-based semantic disorder, and the reacquisition of face identification ability," *Neuropsychological rehabilitation*, vol. 12, no. 1, pp. 1–26, 2002.

[112] E. Mayer, B. Rossion, O. Godefroy, *et al.*, "The behavioral cognitive neurology of stroke," 2007.

[113] J. M. DeGutis, S. Bentin, L. C. Robertson, and M. D'Esposito, "Functional plasticity in ventral temporal cortex following cognitive rehabilitation of a congenital prosopagnosic," *Journal of Cognitive Neuroscience*, vol. 19, no. 11, pp. 1790–1802, 2007.

[114] K. A. Dalrymple, S. Corrow, A. Yonas, and B. Duchaine, "Developmental prosopagnosia in childhood," *Cognitive neuropsychology*, vol. 29, no. 5-6, pp. 393–418, 2012.

[115] J. DeGutis, S. Cohan, D. A. Kahn, G. K. Aguirre, and K. Nakayama, "Facial expression training improves emotion recognition and changes neural tuning in a patient with acquired emotion recognition deficits and prosopagnosia," *Journal of Vision*, vol. 13, no. 9, pp. 993–993, 2013.

[116] J. DeGutis, S. Cohan, and K. Nakayama, "Holistic face training enhances face processing in developmental prosopagnosia," *Brain*, vol. 137, no. 6, pp. 1781–1798, 2014.

[117] S. L. Corrow, J. Davies-Thompson, K. Fletcher, C. Hills, J. C. Corrow, and J. J. Barton, "Training face perception in developmental prosopagnosia through perceptual learning," *Neuropsychologia*, vol. 134, p. 107196, 2019.

[118] J. Davies-Thompson, K. Fletcher, C. Hills, R. Pancaroglu, S. L. Corrow, and J. J. Barton, "Perceptual learning of faces: A rehabilitative study of acquired prosopagnosia," *Journal of Cognitive Neuroscience*, vol. 29, no. 3, pp. 573–591, 2017.

[119] N. Kanwisher, J. McDermott, and M. M. Chun, "The fusiform face area: a module in human extrastriate cortex specialized for face perception," *Journal of neuroscience*, vol. 17, no. 11, pp. 4302–4311, 1997.

[120] I. Gauthier, M. J. Tarr, A. W. Anderson, P. Skudlarski, and J. C. Gore, "Activation of the middle fusiform 'face area' increases with expertise in recognizing novel objects," *Nature Neuroscience*, vol. 2, no. 6, pp. 568–573, 1999.

[121] D. O. Hebb, *The Organization of Behavior: A Neuropsychological Theory*. Wiley, 1949.

[122] E. T. Rolls and M. J. Tovee, "The responses of single neurons in the temporal visual cortical areas of the macaque when more than one stimulus is present in the receptive field," *Experimental Brain Research*, vol. 103, no. 3, pp. 409–420, 1995.

[123] E. H. F. De Haan and B. Duchaine, "Prosopagnosia: A face-specific disorder," *Cognition*, vol. 89, no. 1, pp. 1–25, 2003.

[124] J. e. a. Mann, "Training can enhance facial recognition capabilities," *Journal of Cognitive Neuroscience*, vol. 32, no. 4, pp. 742–755, 2020.

[125] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[126] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[127] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[128] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *BMVC*, vol. 1, p. 6, 2015.

[129] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[130] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1701–1708, 2014.

[131] E. T. Rolls, *The neuronal bases of perceptual and cognitive processes*. Oxford University Press, 2012.

[132] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman, "Building machines that learn and think like people," *Behavioral and Brain Sciences*, vol. 40, 2017.

[133] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[134] P. J. Phillips, A. N. Yates, Y. Hu, C. A. Hahn, E. Noyes, K. Jackson, J. G. Cavazos, G. Jeckeln, R. Ranjan, S. Sankaranarayanan, *et al.*, "Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms," *Proceedings of the National Academy of Sciences*, vol. 115, no. 24, pp. 6171–6176, 2018.

[135] E. Zhou, Z. Cao, and Q. Yin, "Naive-deep face recognition: Touching the limit of lfw benchmark or not?," *arXiv preprint arXiv:1501.04690*, 2015.

[136] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of big data*, vol. 6, no. 1, pp. 1–48, 2019.

[137] H. Lee, S.-H. Park, J.-H. Yoo, S.-H. Jung, and J.-H. Huh, "Face recognition at a distance for a stand-alone access control system," *Sensors*, vol. 20, no. 3, p. 785, 2020.

[138] R. J. Itier, M. Latinus, and M. J. Taylor, "Face, eye and object early processing: what is the face specificity?," *Neuroimage*, vol. 29, no. 2, pp. 667–676, 2006.

[139] B. De Gelder, A.-C. Bachoud-Lévi, and J.-D. Degos, "Inversion superiority in visual agnosia may be common to a variety of orientation polarised objects besides faces," *Vision research*, vol. 38, no. 18, pp. 2855–2861, 1998.

[140] S. Baron-Cohen, S. Wheelwright, Jolliffe, and Therese, "Is there a" language of the eyes"? evidence from normal adults, and adults with autism or asperger syndrome," *Visual cognition*, vol. 4, no. 3, pp. 311–331, 1997.

[141] S. Mann, Z. Pan, Y. Tao, A. Gao, X. Tao, D. E. Garcia, D. Shi, and G. Kannan, "Face recognition and rehabilitation: a wearable assistive and training system for prosopagnosia," in *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 731–737, IEEE, 2020.

[142] Y. Li and S. Lyu, "Exposing deepfake videos by detecting face warping artifacts," *arXiv preprint arXiv:1811.00656*, 2018.

[143] L. Chai, D. Bau, S.-N. Lim, and P. Isola, "What makes fake images detectable? understanding properties that generalize," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16*, pp. 103–120, Springer, 2020.

[144] P. Yu, Z. Xia, J. Fei, and Y. Lu, "A survey on deepfake video detection," *Iet Biometrics*, vol. 10, no. 6, pp. 607–624, 2021.

[145] G. An, M. Akiba, K. Omodaka, T. Nakazawa, and H. Yokota, "Hierarchical deep learning models using transfer learning for disease detection and classification based on small number of medical images," *Scientific reports*, vol. 11, no. 1, p. 4250, 2021.

[146] J. Latif, C. Xiao, A. Imran, and S. Tu, "Medical imaging using machine learning and deep learning algorithms: a review," in *2019 2nd International conference on computing, mathematics and engineering technologies (iCoMET)*, pp. 1–5, IEEE, 2019.

[147] N. Waqas, S. I. Safie, K. A. Kadir, S. Khan, and M. H. K. Khel, "Deepfake image synthesis for data augmentation," *IEEE Access*, vol. 10, pp. 80847–80857, 2022.

[148] V. Thambawita, J. L. Isaksen, S. A. Hicks, J. Ghouse, G. Ahlberg, A. Linneberg, N. Grarup, C. Ellervik, M. S. Olesen, T. Hansen, *et al.*, "Deepfake electrocardiograms using generative adversarial networks are the beginning of the end for privacy issues in medicine," *Scientific reports*, vol. 11, no. 1, p. 21896, 2021.

[149] B. Zhu, H. Fang, Y. Sui, and L. Li, "Deepfakes for medical video de-identification: Privacy protection and diagnostic information preservation," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 414–420, 2020.

[150] R. Budhiraja, M. Kumar, M. Das, A. S. Bafila, and S. Singh, "Medifaked: Medical deepfake detection using convolutional reservoir networks," in *2022 IEEE Global Conference on Computing, Power and Communication Technologies (GlobConPT)*, pp. 1–6, IEEE, 2022.

[151] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, pp. 13001–13008, 2020.

[152] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," *arXiv preprint arXiv:1708.04552*, 2017.

[153] S. A. Khan and D.-T. Dang-Nguyen, "Hybrid transformer network for deepfake detection," in *Proceedings of the 19th International Conference on Content-based Multimedia Indexing*, pp. 8–14, 2022.

[154] L. Chen, Y. Zhang, Y. Song, L. Liu, and J. Wang, "Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18710–18719, 2022.

[155] D. Ko, S. Lee, J. Park, S. Shin, D. Hong, and S. S. Woo, "Deepfake detection for facial images with facemasks," *arXiv preprint arXiv:2202.11359*, 2022.

[156] S. Das, S. Seferbekov, A. Datta, M. S. Islam, and M. R. Amin, "Towards solving the deepfake problem: An analysis on improving deepfake detection using dynamic face augmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3776–3785, 2021.

[157] Google, "Mediapipe face mesh documentation." https://github.com/google/mediapipe/blob/master/docs/solutions/face_mesh.md, 2022. Accessed on 2023-10-13.

[158] 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*, pp. 1–11, 2019.

[159] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-df: A large-scale challenging dataset for deepfake forensics," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3207–3216, 2020.

[160] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. C. Ferrer, "The deepfake detection challenge (dfdc) dataset," *arXiv preprint arXiv:2006.07397*, 2020.

[161] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1251–1258, 2017.

[162] X. Huang, G. Zhao, W. Zheng, and M. Pietikäinen, "Towards a dynamic expression recognition system under facial occlusion," *Pattern Recognition Letters*, vol. 33, no. 16, pp. 2181–2191, 2012.

[163] Y. Nirkin, L. Wolf, Y. Keller, and T. Hassner, "Deepfake detection based on discrepancies between faces and their context," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 6111–6121, 2021.

[164] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

[165] 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)*, pp. 83–92, IEEE, 2019.

[166] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "Mesonet: a compact facial video forgery detection network," in *2018 IEEE international workshop on information forensics and security (WIFS)*, pp. 1–7, IEEE, 2018.

[167] H. Lin, W. Huang, W. Luo, and W. Lu, "Deepfake detection with multi-scale convolution and vision transformer," *Digital Signal Processing*, vol. 134, p. 103895, 2023.

[168] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo, "Face x-ray for more general face forgery detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5001–5010, 2020.

[169] I. Masi, A. Killekar, R. M. Mascarenhas, S. P. Gurudatt, and W. AbdAlmageed, "Two-branch recurrent network for isolating deepfakes in videos," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pp. 667–684, Springer, 2020.

[170] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic, "Lips don't lie: A generalisable and robust approach to face forgery detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5039–5049, 2021.

[171] A. Ismail, M. Elpeltagy, M. S. Zaki, and K. Eldahshan, "A new deep learning-based methodology for video deepfake detection using xgboost," *Sensors*, vol. 21, no. 16, p. 5413, 2021.

[172] Q. Liao, H. Chai, H. Han, X. Zhang, X. Wang, W. Xia, and Y. Ding, "An integrated multi-task model for fake news detection," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 11, pp. 5154–5165, 2021.

[173] Y. Ding, W. Zhang, X. Zhou, Q. Liao, Q. Luo, and L. M. Ni, "Fraudtrip: Taxi fraudulent trip detection from corresponding trajectories," *IEEE Internet of Things Journal*, vol. 8, no. 16, pp. 12505–12517, 2020.

[174] H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang, and N. Yu, "Multi-attentional deepfake detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2185–2194, 2021.

[175] S. Wang, J. Zhou, and H. Yang, "Improved face forensics with high-quality synthetic face data," in *2019 IEEE International Conference on Computer Vision (ICCV)*, pp. 1195–1204, IEEE, 2019.

[176] J. Choi and Y. Kim, "Colorful cutout: Enhancing image data augmentation with curriculum learning," *ArXiv*, 2024.

[177] 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)*, pp. 5012–5019, IEEE, 2021.