Anomaly Detection in Face Anti-spoofing: Algorithms, Training Set Construction, and Bias Analysis.

LATIFAH ABDULLAH A ABDUH

Anomaly Detection in Face Anti-spoofing: Algorithms, Training Set Construction, and Bias Analysis

Latifah Abdullah A.Abduh

A Thesis presented for the degree of
Doctor of Philosophy

Department of Computer Science
Durham University
United Kingdom
May 2023

Abstract

Face recognition is a mature and trustworthy method for identifying individuals. Thanks to the availability of high-definition cameras and accompanying devices, this particular biometric recognition modality is widely regarded as the fastest and least obtrusive option. Despite advancements in face recognition systems, it has been discovered that successful spoofing attempts are still possible. Various anti-spoofing algorithms, also known in the literature as liveness detection tests and presentation attack detection algorithms, have been devised to counteract such attacks.

The first contribution of this research is to demonstrate the effectiveness of certain simple and direct spoofing attacks. Our approach involves utilizing ResNet50, a highly reliable deep neural network, as a binary classification method. We assess its performance by subjecting it to adversarial attacks that involve manipulating the saturation component of imposter images. We have found that it is particularly vulnerable to spoofing attacks that employ processed imposter images. To the best of our knowledge, this study represents the pioneering exploration of adversarial attacks on deep neural networks within the realm of face anti-spoofing detection. In addition, we conducted an experiment that revealed the potential of the proposed adversarial attack to be converted into a direct presentation attack.

In a second contribution, we propose an alternative approach incorporating in-the-wild images and non-specialised databases into anomaly detection to improve the face anti-spoofing algorithm's performance on unseen databases. We developed a method for detecting anomalies in face anti-spoofing by employing a convolutional autoencoder. We assessed its effectiveness using the NUAA database, which had not been previously utilized in the training. Our results indicated improved performance when incorporating in-the-wild face images and face data from non-specialized databases into the training dataset.

Transformers are emerging as the new gold standard in various computer vision applications and have already been used in face anti-spoofing, demonstrating competitive performance. In a third contribution, we propose a network with the ViT transformer and ResNet18 as the backbone for anomaly detection in face anti-

spoofing with a decoder as the head. Then, we validate various anomaly detectors to compare the results with our proposed method. Also, using the ViT with MLP as a binary classifier baseline and compare it with our model. Our comprehensive testing and evaluation have demonstrated that this proposed approach competes admirably as a method for detecting anomalies in the domain of face anti-spoofing.

Finally, there are only a few papers that specifically address the issue of racial bias in anti-spoofing. As a fourth contribution, we present a systematic study of race bias in face anti-spoofing with three key characteristics: the focus is on analysing potential bias in bona fide errors, where significant ethical and legal issues lie; analyses of various stages of the classification process, and treating the value of the threshold that determines the classifier's operating point on the ROC curve as a user-defined variable. We do not assume it is fixed by the vendor of the biometric verification system through a black-box process. To the best of our knowledge, this is the first investigation into racial bias within the face anti-spoofing domain that employs anomaly detection techniques while also incorporating a non-specialized database for analysis. Our results show that racial bias in face anti-spoofing is influenced by factors beyond mean response values, such as different variances, bimodality, and outliers.

Overall, this thesis contributes to the ongoing development of anti-spoofing techniques and investigates some important issues regarding the potential for bias in these systems.

Keywords: Face spoofing attacks, Face anti-spoofing, Adversarial attack, Anomaly Detection, Racial Bias.

Declaration

The work in this thesis is based on research carried out at the Department of Computer Science, Durham University, United Kingdom. No part of this thesis has been submitted elsewhere for any other degree or qualification and it is all my own work unless referenced to the contrary in the text.

Acknowledgements

Praise be to Allah (SWT), I offer my sincerest gratitude to Allah (SWT) for the opportunity to embark on this educational pursuit and for granting me the capability to pursue knowledge. I am immensely grateful for His divine guidance, which has been the guiding light leading me towards academic success and personal growth.

I would like to express my deepest gratitude to my supervisor, *Dr. Ioannis Ivrissimtzis*, for his invaluable guidance, support, and expertise throughout this research. His insightful feedback and encouragement have been instrumental in shaping the direction of this study. Also, I would like to thank *Dr.Frederick Li*, for his support as a second supervisor.

I would like to convey my sincere appreciation to *Dr. Luma Omar* for her invaluable cooperation. Her contribution has been instrumental in achieving our shared objectives, and I am truly grateful for her dedication and expertise. It has been a pleasure working with such a talented and committed colleague.

My father, *Dr. Abdullah Abduh*, thank you for instilling a strong work ethic and always pushing me to reach my full potential. Your wise counsel, motivation, and unwavering belief in my abilities have been invaluable.

My mother, *Mrs. Naimah Al-adresey*, I appreciate your unwavering support. Your endless patience, belief in my abilities, and countless sacrifices have inspired me to strive for excellence.

My second mother, *Mrs. Fyzah Al-zahrani*, I am grateful for the genuine care and affection you have shown me. Your guidance and encouragement have motivated me throughout my educational pursuits.

My dear husband, *Dr. Hatem Nojoum*, thank you for being my partner in this journey and my biggest cheerleader. Your constant love, patience, and belief in me have given me the strength and motivation to overcome challenges and persevere.

My beautiful daughters, *Leen & Loreen*, thank you for being my inspiration and motivation. Your love, laughter, and joy have filled my life with purpose and happiness.

My beloved brothers and sisters, *Abdullaziz, Ghadi, Modhi, Abdulrahman, Howazen*

and Muhammed , I want to express my heartfelt gratitude and appreciation to each and every one of you.

I want to extend my heartfelt appreciation and gratitude to my beloved family-in-law. Your love, acceptance, and support have been integral to my life.

I would like to thank my dear friends, *Omaima, Ruba, Jood, Layla and Muna* and all my friends in Newcastle, Durham and Saudi Arabia for their unwavering support and friendship throughout my thesis journey.

Contents

Abstract ii

Declaration iv

Acknowledgements v

List of Figures xii

List of Tables xvi

Dedication xix

Publications xx

1 Introduction 1

1.1 Motivation 4

1.2 Problem Statement 5

1.3 Research Challenges 6

1.4 Research Questions and Objectives 8

1.5 Research Methodology 10

1.6 Contributions 11

1.7 Thesis Structure 14

2 Background on Biometrics **15**

2.1 Biometric Identification 16

2.1.1 Biometric Traits 16

2.2 Face Recognition Systems 17

2.3 Vulnerabilities of Face Recognition Systems 19

2.4 Types of Presentation Attacks 21

2.5 Face Anti-spoofing Methods 24

2.5.1 Software-based Methods 25

2.5.2 Face Anti-spoofing Methods based on Special Hardware 35

2.6 Databases 36

2.6.1 Face Anti-spoofing Databases 36

2.6.2 Face Recognition Databases 43

2.6.3 Face Detection 45

2.7 Evaluation Metrics 47

2.8 Evaluation protocols 49

2.8.1 Intra-dataset Intra-Type Protocol 49

2.8.2 Cross-Dataset Intra-Type Protocol 50

2.9 Conclusion 50

3 Face Anti-spoofing Methods Based on Anomaly Detection **52**

3.1 Introduction 52

3.2 Anomaly detection in Face Anti-spoofing 55

3.2.1 Handcrafted-Features with One Class Classification 55

3.2.2 Deep Learning with One Class-classification 56

3.3 Summary of The Main Findings 64

3.4 Conclusion 66

4 Colour Processing in Adversarial Attacks on Face Anti-spoofing **67**

4.1 Introduction 68

4.2 Related work 70

4.2.1 Adversarial Attacks 70

4.3 Methods 72

4.3.1 ResNet . . . 72
4.3.2 Color Manipulation Technique . . . 73
4.4 Experimental setup . . . 73
4.4.1 Implementation and training . . . 73
4.4.2 Validation . . . 74
4.5 Results and Discussion . . . 74
4.5.1 Adversarial Attack . . . 74
4.5.2 Presentation Attack . . . 77
4.6 Conclusions . . . 79

5 Training Dataset Construction for Anomaly Detection in Face Anti-spoofing 80
5.1 Introduction . . . 80
5.2 Related work . . . 83
5.2.1 Autoencoders for Face Anti-spoofing . . . 83
5.3 Methods . . . 83
5.3.1 Anomaly scores . . . 85
5.4 Experiment . . . 86
5.4.1 Convolutional Autoencoder . . . 86
5.4.2 Training, validation, and test datasets . . . 87
5.5 Results and Discussion . . . 90
5.6 Conclusions . . . 93

6 Anoformer PAD: Anomaly Detection with Transformers for Face Anti-spoofing 94
6.1 Introduction . . . 94
6.2 Related Work . . . 97
6.2.1 ViT Transformers for face anti-spoofing . . . 97
6.3 Methods . . . 98
6.3.1 Transformer . . . 98
6.3.2 Frechet Inception Distance (FID) . . . 99
6.3.3 One Class Classifiers . . . 100

6.4 The Anoformer . . . 100
6.4.1 Architecture . . . 101
6.4.2 Implementation . . . 103
6.5 Results and Discussion . . . 104
6.5.1 Databases . . . 104
6.5.2 Evaluation Metrics . . . 104
6.5.3 Anoformer validation . . . 105
6.5.4 Performance evaluation . . . 107
6.6 Conclusion . . . 109

7 Race Bias Analysis of Bona Fide Errors in face anti-spoofing 111
7.1 Introduction . . . 112
7.2 Related Work . . . 114
7.2.1 Bias in machine learning . . . 114
7.2.2 Bias in Presentation Attack Detection . . . 116
7.2.3 Databases . . . 116
7.3 Methods . . . 117
7.3.1 Vector Quantized Variational Autoencoder . . . 117
7.3.2 Support Vector Machine . . . 119
7.3.3 Statistic Analysis . . . 120
7.4 Experimental setup . . . 121
7.4.1 The VQ-VAE network . . . 121
7.4.2 Data preparation . . . 122
7.4.3 Network validation . . . 123
7.5 Bias analysis on SiW . . . 124
7.5.1 Statistical analysis of the binary outcomes . . . 125
7.5.2 Statistical analysis of the scalar responses . . . 126
7.5.3 Discrete latent space . . . 131
7.6 Bias analysis on RFW . . . 132
7.6.1 Statistical analysis of the scalar responses . . . 134
7.6.2 Discrete latent space . . . 138
7.7 Conclusion . . . 139

8 Conclusion **140**

8.1 Introduction 140

8.2 Research contributions 141

8.3 Limitations 143

8.4 Privacy Implications of Cloud-Based Face Anti-Spoofing 143

8.5 Future Work 144

List of Figures

1.1 Physiological and behavioural biometric modalities are the two main categories of biometric techniques. 2
1.2 Genuine and imposter access to face recognition 4
1.3 From left to right: the first images are imposters from the database, and the processed imposter images with saturation values 63 and 223, respectively. 12
1.4 Some examples from the harvested wild images. 12
1.5 The imposter-cut-photo from Casia-FAS DB (left) and the visualization of its attention (right) 13

2.1 From 1 to 8, there are eight potential attack places. [1] 20
2.2 Illegitimate access types to face recognition systems: direct and indirect. 21
2.3 Types of face presentation attacks. 22
2.4 Our research taxonomy of face anti-spoofing methods 25
2.5 An illustration of the method proposed by Matta et al. [2] 26
2.6 Some examples of genuine faces and print and video representations are presented in RGB, grayscale, and Y CbCr colour spaces, respectively [3]. 27
2.7 The proposed method by Atom et al [4]. 29

2.8 Imposters examples from NUAA DB [5]. 38
2.9 Replay-Attack samples from a controlled setting (top row, left to right) and an uncontrolled environment (bottom row, left to right) comprise clients, printed photos, mobile phone and tablet photo attacks [6]. 39
2.10 Examples of live and spoof videos in the SiW DB top row and bottom row, respectively [7]. 41
2.11 Examples from SCface DB. 45
2.12 Examples from each race group in RFW DB. 45

3.1 (a) (Binary classification) genuine and known attacks are used in training. (b) (One class classification) Only genuine images are available during training [8]. 54
3.2 The proposed method in [9] consists of a generator and a discriminator trained to understand the underlying structure in the live faces data. 64

4.1 (a) a standard CNN (b) a Residual Block in ResNet. 72
4.2 ResNet-50 neural network architecture with adding the custom classifier layers. 74
4.3 Our proposed method to generate the Adversarial Attack. 75
4.4 Saturation linearly scaled by a constant α and capped to 255. From left to right: $\alpha = 0, 0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75$. 76
4.5 Fixed saturation values s. From left to right: $s = 31, 63, 95, 127, 159, 191, 223, 255$. 77
4.6 Presentation Attack. 78
4.7 Paper print: for each pair of images, the left is a photo of the original client and the right is a photo of the processed client. 78
4.8 LCD: for each pair of images, the left is a photo of the original client and the right a photo of the processed client. 79
4.9 iPhone: for each pair of images, the left is a photo of the original client, and the right is a photo of the processed client. 79

5.1 Example of the real and reconstructed image with their MSE for both client and imposter classes. 85

5.2 On the left, client images alongside their corresponding reconstructed images generated by the CAE, along with the associated MSE values. On the right, impostor images are presented alongside their reconstructed counterparts, also accompanied by their respective MSE values. 86
5.3 The architecture of the proposed convolutional autoencoder. 87
5.4 The first row are Examples of wild images and the second row are examples from face recognition DBs. 90
5.5 ROC curves corresponding to classifier/training dataset combinations, tested on Replay-Attack (top) and NUAA (bottom). 91

6.1 Architecture of the Vision transformer from [10]. 98
6.2 Architecture of the Anoformer. 101
6.3 Visualisation of input faces from the Replay Attack database, and the attention maps of several ViT layers. **Top two rows:** a bona fide face. **Bottom two rows:** an imposter face. 110

7.1 The architecture of the proposed VQ-VAE. 122
7.2 **Top:** The responses of the RA trained network on each race of the SiW testset. From left to right: African, Asian, Caucasian, Indian. **Bottom:** The responses of the SiW trained network. 122
7.3 The bias analysis process. The binary outcome analysis is shown in purple, the scalar responses analysis in blue, and the latent space analysis in orange. 125
7.4 For each pair of races, graphs of the p-value as a function of the threshold. The classifier was trained on RA. 126
7.5 For each pair of races, graphs of the p-value as a function of the threshold. The classifier was trained on SiW. 127
7.6 For each pair of races in SiW, the histogram of the responses on bona-fide images. The classifier was trained on RA. 128
7.7 For each pair of races in SiW, the histogram of the responses of bona-fide images. The classifier was trained on SiW. 129

7.8 For each pair of races, graphs of the p-value as a function of the threshold. The classifier was trained on RA. 135
7.9 For each pair of races, graphs of the p-value as a function of the threshold. The classifier was trained on SiW. 136
7.10 For each pair of races in RFW, the histogram of the responses of bona-fide images. The classifier was trained on RA. 137
7.11 For each pair of races in RFW, the histogram of the responses of bona-fide images. The classifier was trained on SiW. 138

List of Tables

2.1 A summary of publicly available datasets was used in anomaly detection methods for face anti-spoofing. This table represents the year, the number of subjects, PA types, ethnicity, pose, expression and PAI modality in these DBs. In the column 'PA type', 'P, 'R', and 'M' denote 'photo' and 'replay-video' and 'masks', respectively. 37
2.2 Evaluation metrics and their equations. 51

3.1 HTER (↓) results of client-specific methods for Replay-Mobile and Rose-Youtu DB 61
3.2 AUC (%) and HTER results of intra-database test on the Replay-Attack DB using Deep learning methods. 63
3.3 Comparison of face anti-spoofing methods on cross-database testing CASIA-MFSD (C)and Replay-Attack datasets (RA). The top four rows of HTER results correspond to binary methods, and the bottom two rows are the one-class methods. 65

4.1 TPR T_p and average loss L for various values of α. The bold numbers corresponds to the original images. 76
4.2 TPR Tp and loss L for various fixed values of s 77
4.3 TPR and (loss) for all 6 types of presentation attacks. 78

5.1 Description and size of the training datasets. 89
5.2 The AUC values corresponding to the ROC curves shown in Figure 5.5. 91
5.3 HTERs computed on the operating point corresponding to the EER of the validation set. The last column shows the range of HTERs reported in [11]. 92

6.1 ViT + ResNet backbone with various one-class classifiers tested on RA. 106
6.2 ViT + ResNet backbone with various one-class classifiers tested on SiW. 106
6.3 Ablation study for the Anoformer backbone. 107
6.4 Performance comparison against [8]. 107
6.5 Comparison against ViT feature vector + two-class trained MLP, in intra- and cross-database testing. We report AUC values on each presentation attack species separately. 108
6.6 Intra- and cross-database testing of the Anoformer with threshold-specific metrics. 109

7.1 Sizes of training and test datasets. The test sets are equally split between clients and imposters. The SiW test set, consists of 400 samples from each race. 123
7.2 Error rates computed under the testing protocol. 124
7.3 HTERs computed for each race separately on the SiW testset. 124
7.4 p-values of the chi-squared tests for the thresholds in Section 7.4.3 . . 125
7.5 p-values of the Mann–Whitney U test on each pair of races. 128
7.6 SiW testset: means, standard deviations, and Hartigan's dip values for the responses of the RA and SiW trained classifiers. 131
7.7 AUC values for an SVM trained on the VQ-VAE's latent space encodings of the images. SiW testset. 132
7.8 Threshold values corresponding to some predetermined TPR values. . 133
7.9 p-values of the chi-squared tests for the thresholds shown in Table 7.8. 134
7.10 p-values of the Mann–Whitney U test on each pair of races. 136

7.11 RFW testset: means, standard deviations, and Hartigan's dip values for the responses of the RA and SiW trained classifiers. 137
7.12 AUC values for an SVM trained on the VQ-VAE's latent space encodings of the images. RFW testset. 139

Dedication

To my beloved grandmother, she was a source of strength, wisdom, and inspiration. Her unconditional love and belief in my abilities fueled my determination to pursue my academic goals. She gave me endless encouragement and taught me the value of hard work, perseverance, and dedication. "Rahma Hurasani", may ALLAH rest her soul in an eternal piece and a place in the highest ranks of Paradise.

Publications

The contents of this thesis are based on the results of the following papers.

Publications

- Abduh, L., and Ivrissimtzis, I. Colour processing in adversarial attacks on face liveness systems. In CGVC 2019, pp. 149-152 (Chapter 4)

- Abduh, L., and Ivrissimtzis, I. Training dataset construction for anomaly detection in face anti-spoofing. In CGVC 2021, (Chapter 5)

- Abduh, L., Omer, L., and Ivrissimtzis, I. Anomaly Detection with Transformers in Face Anti-spoofing. In WSCG 2023, (Chapter 6)

- Abduh, L., and Ivrissimtzis, I. Race Bias Analysis of Bona Fide Errors in Face anti-spoofing (Chapter 7)

In addition to the above, the following paper were submitted

- Abduh, L., Omer, L., and Ivrissimtzis, I. Anomaly Detection in Face Anti-Spoofing : Survey. In IET, (Chapter 3)

CHAPTER 1

Introduction

Personal identification plays an essential role in many of our daily activities. The most traditional personal identification techniques are knowledge-based and token-based, such as passwords and a passport, respectively [12]. However, these methods are vulnerable to theft or misuse, which limits their effectiveness. Moreover, they cannot provide crucial functions such as non-repudiation or detecting multiple instances [13]. Biometrics-based user verification systems extract unique biological characteristics from individuals and use statistical analysis to verify their identity. Using biometric characteristics for identification is becoming more popular due to their uniqueness.

Over the past few decades, biometric security has rapidly grown in popularity and can compete against traditional methods. Biometric characteristics are any measurable differentiating traits utilized for biometric identification. As shown in Figure 1.1, physiological and behavioural biometrics are the two primary types of biometric traits. Physiological biometrics, such as fingerprint, facial, and iris recognition, rely on measurements of human body parts [14]. Behavioural biometrics quantify human actions and include gesture, key stroking, gait, and signature recognition [13, 15].

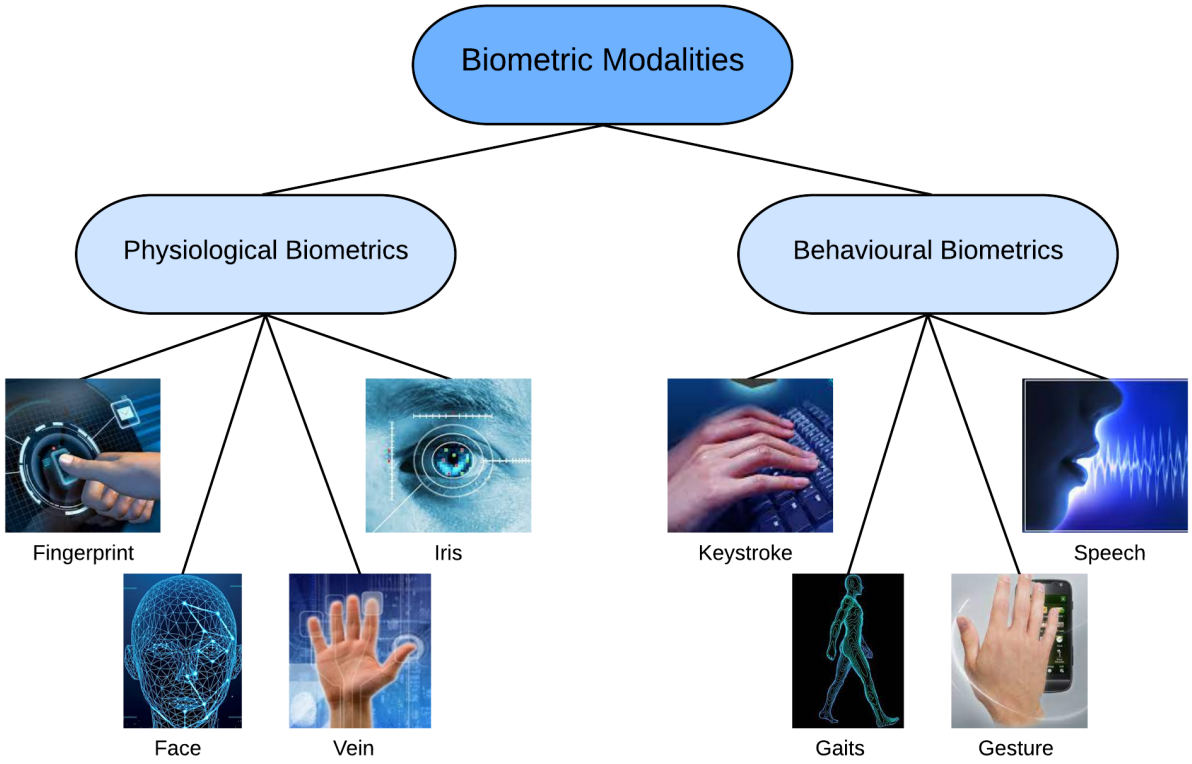

Figure 1.1: Physiological and behavioural biometric modalities are the two main categories of biometric techniques.

One of the main advantages of biometric authentication is its ability to provide strong and reliable security. Unlike traditional authentication methods, such as passwords or PINs, biometric authentication relies on unique physical or behavioral traits that are difficult to replicate or steal. This makes biometric authentication more secure and less vulnerable to fraud or hacking. Additionally, biometric authentication offers convenience and ease of use for users. With biometric authentication, users do not need to remember and manage multiple passwords or tokens, which can be a hassle and a source of frustration. Instead, they can use their own body to authenticate themselves, which is faster, more efficient, and more natural [16].

One of the most widely used biometric authentication methods is face recognition, which provides easy and secure access to various systems ranging from personal mobile phones and laptops to online banking and electronic airport security. Face recognition has drawn interest recently from both the industrial and academic worlds [17]. The use of face biometrics has many advantages: it is natural, fast, easy to use, reliable, less human invasive, uses non-intrusive data and employs low-cost

sensors. It has already found a wide range of uses as a convenient and affordable method of personal identification, from security-critical applications like passport control at airport gates to consumer-level applications like automatically logging into a laptop or smartphone.

Despite the many advances in face recognition technology, it is still vulnerable to various types of spoofing attacks, especially presentation attacks, also known as direct attacks. These attacks aim to bypass an authentication system by displaying images or replaying videos in front of the system's camera or by wearing a 3D mask of the authorized user [18]. Figure 1.2 illustrates the access process for a face recognition system, involving two distinct categories of individuals: genuine users and imposters.

Even though many studies have been published on face recognition systems in recent years, presentation attacks are still considered an extremely serious threat [6]. Any potential imposter can easily obtain images or videos of a genuine user from the Internet and social media.

This thesis focuses on the security risk of face biometric systems, specifically face spoofing attacks. These attacks involve a malicious user presenting a counterfeit or manipulated face image to the system in an attempt to falsely claim the identity of another user. Face spoofing attacks pose a serious risk to the success of the widespread implementation of face biometric systems, particularly in circumstances where human supervision is impractical or impossible.

The rest of this Chapter is organised as follows: In Section 1.1, we present the motivation behind this thesis. Then, in Section 1.2, we discuss the problem statement of this research, followed by the research challenges in Section 1.3. The research questions and objectives of this thesis are formalized in Section 1.4. Section 1.5 provides a summary of the methodologies used in this thesis. In Section 1.6, we present the main contributions of this thesis. Finally, the thesis structure is outlined in Section 1.7.

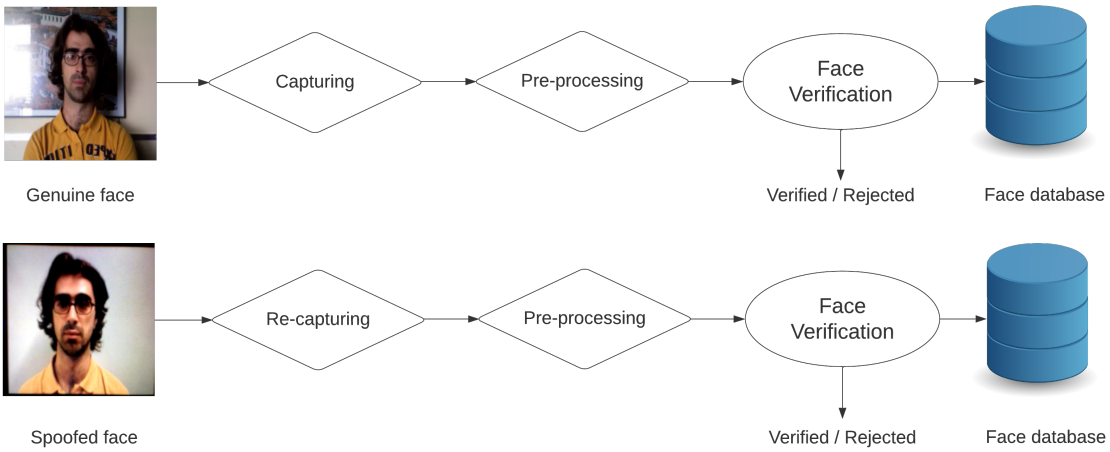

Figure 1.2: Genuine and imposter access to face recognition

1.1 Motivation

Biometrics are unique but not always private, making it possible to copy a biometric trait, which has become even easier due to the increasing availability of biometric data online. Once an attacker successfully copies a biometric trait, they can gain access to the system and prevent the legitimate user from using that trait again. It is crucial to note that spoofing attacks on biometrics are not merely theoretical concerns but practical problems that significantly impact the widespread adoption of biometric authentication. The vulnerability of commercial fingerprint recognition equipment to spoofing attacks using gummy fingers was demonstrated, highlighting the need for enhanced security measures [19]. In response, security enthusiasts have been experimenting with new biometric authentication systems. For instance, at the Black Hat Security conference, [20] were able to deceive multiple laptop face authentication systems using fake facial images.

Face recognition technology has gained significant popularity and widespread adoption across various domains. In a study by [21], an experiment was conducted to evaluate the effectiveness of several user verification systems that rely on face recognition technology. The objective was to assess the resilience of these systems against simple identity spoofing methods, such as using ID cards or online photos of users. Four popular user verification tools were tested: Luxand Blink on Windows [22], Keyemon [23], FaceLock on Android [24], and Android Face Unlock [25].

The experiment results indicated that all tested systems were vulnerable to even basic attacks, suggesting that the technology is not yet suitable for applications prioritizing security over user convenience. These findings emphasize the need for further face recognition technology advancements to enhance its robustness against identity spoofing techniques.

Face anti-spoofing (FAS) is an active research area within computer vision, and a growing number of publications have been dedicated to it in recent years. Typically, FAS algorithms are evaluated using databases that consist of two types of images: client images, which are client's photos of real individuals, and imposter images, which are spoofing photos created by modifying client images. Constructing such databases poses challenges due to the various sources of variation in spoofing attacks. The success of a spoofing attack can be influenced by factors such as the use of a printed photo or an electronic display, the quality of the paper and printer, the dimensions of the photo, and how it is presented to the camera. These factors impact both the success of the attack and the perceived effectiveness of the anti-spoofing algorithm.

FAS aims to detect whether the presented face is real or a fake representation of the claimed identity. The goal is to reject the fake representations and only accept the real ones. FAS can be performed either in a passive manner, where the system only relies on the face image or video, or in an active manner, where the system is equipped with additional sensors that provide further information about the face, such as depth or thermal information. For the last few years, a significant amount of headway has been made toward developing robust algorithms for detecting face spoofing [26]. Researchers have dedicated considerable efforts to this area, advancing the capabilities of FAS in identifying and mitigating face spoofing attacks.

1.2 Problem Statement

Current FAS methods often rely on handcrafted features or deep learning models trained on large datasets of both real and fake face images. However, these methods may not be robust to new, unseen types of attacks or variations in lighting conditions

and may be vulnerable to adversarial attacks.

Anomaly detection is a promising approach to improving FAS, as it can help detect previously unseen or novel spoofing attacks that traditional methods may not be able to identify [27]. By combining the strengths of FAS and anomaly detection, researchers can develop more robust and accurate methods for detecting spoofing attempts, enhancing the security and reliability of facial recognition technology.

Therefore, research on anomaly detection methods for FAS has the potential to significantly improve the effectiveness and efficiency of FAS techniques, with important implications for biometric authentication systems in a range of applications, from mobile devices to border security.

By exploring the effectiveness of one-class classification methods for FAS using anomaly detection techniques, this thesis can help develop a more reliable and scalable method for detecting various spoofing attacks under different conditions, ultimately improving the security and trustworthiness of face recognition systems.

1.3 Research Challenges

Despite significant recent progress in the field [28], several challenges persist, including the need for transparency and robustness in face anti-spoofing (FAS) technology. This challenge is exemplified by the research quest for explainability, focusing on the development of AI models that can distinguish between genuine and impostor faces while providing clear, interpretable explanations for their decisions. This involves crafting face representations that capture crucial facial features while maintaining understandability. It also entails the formulation of effective criteria for detecting anomalies and fraudulent activities in face recognition systems and providing insights into the success or failure of specific attacks. Addressing this challenge is pivotal for instilling trust in AI systems and bolstering the security and reliability of facial recognition technology.

Another hurdle faced by FAS technology, particularly those leveraging deep learning approaches, arises from the absence of comprehensive benchmark datasets [29]. The creation of such datasets is a challenging and time-consuming endeavor,

leading to the limited availability of comprehensive face spoofing datasets. Consequently, these methods are susceptible to overfitting, struggling to handle data variations unencountered during training. Deep learning approaches, despite their widespread adoption in FAS research [30, 31], have primarily been evaluated in restricted test scenarios, underlining the urgency for further evaluation and advancement.

Furthermore, existing FAS datasets are often collected under controlled conditions [6]. These conditions involve meticulous control and standardization of factors like lighting, camera angles, and subject placement to ensure consistent data collection. However, FAS systems are intended for real-world deployment, where unpredictable factors like lighting, pose, and occlusions are beyond control. These systems are usually trained to detect specific types of attacks, potentially failing to identify other attack variants, as indicated in [32].

To surmount these limitations, anomaly detection methods have emerged as an alternative approach in FAS by identifying data samples that deviate from normal patterns, often referred to as anomalies or outliers [33, 34]. These methods do not rely on labeled data and are less prone to overfitting compared to deep learning approaches. Typically trained on datasets comprising only client images representing expected normal behavior, anomaly detection methods are gaining traction. However, ongoing research endeavors aim to enhance the generalization capabilities of such FAS methods, especially concerning novel attack types, and to develop effective feature representations to bolster FAS robustness.

Anomaly detection methods offer a promising avenue for FAS research, as they overcome the limitations of deep learning techniques, reducing the need for labeled data and minimizing susceptibility to overfitting. By advancing suitable face representations and deploying effective anomaly detection techniques, FAS can be further fortified, providing a robust and reliable solution for detecting face spoofing.

1.4 Research Questions and Objectives

Because the biometrics community is becoming more interested in the problem of spoofing, the number of publications in the field has grown considerably in recent years. The aim of this thesis is to conduct original research aimed at answering the following **research questions**:

1. Creating enhanced presentation attacks.
    (a) Certain image processing operations on the test set can propose an enhanced adversarial attack. By how much does the accuracy of the FAS algorithms drop?
    (b) Can a proposed adversarial attack on the images of the test set be converted into a direct physical presentation attack?
2. Construction of training sets for anomaly detection.
    (a) To which extent will the performance of the anomaly detection method be enhanced if, in the training set, we include images from in-the-wild and images from non-specialized face databases?
3. Proposing a novel model for FAS based on anomaly detection.
    (a) Can a deep generative model efficiently perform Anomaly Detection and solve the generalization problem in the context of face anti-spoofing?
4. Racial bias in anomaly detection face anti-spoofing.
    (a) How can we investigate racial bias in the face anti-spoofing anomaly model?

The **objectives** of this thesis are to assess the security risks associated with face spoofing attacks, and to develop effective countermeasures for mitigating these risks:

Chapter 3: Face Anti-spoofing Methods Based on Anomaly Detection Survey.

1. To create a comprehensive survey of all current FAS methods based on anomaly detection.

Chapter 4: Colour Processing in Adversarial Attacks on Face Anti-spoofing.

1. To develop methods for enhancing the efficiency of presentation attacks using an image processing technique.
2. To study adversarial attacks on liveness tests based on deep neural networks.
3. To verify that a digital adversarial attack on the test database of the classifier can be converted into a physical attack.

Chapter 5: Training Dataset Construction for Anomaly Detection in Face Anti-spoofing.

1. To prepare and aggregate images from in-the-wild harvested online, and images from facial databases, and use them as a training set for anomaly detection algorithms.
2. To evaluate an Anomaly detection method for FAS based on a convolutional autoencoder that only requires RGB images for input.

Chapter 6: Anoformer PAD: Anomaly Detection with Transformers for Face Anti-spoofing.

1. To develop a robust Anomaly detection algorithm using a transformer with ResNet.
2. To evaluate different one-class classifiers with this backbone (transformer and ResNet).

Chapter 7: Race Bias Analysis of Bona Fide Errors in Face Anti-spoofing.

1. To evaluate the fairness of generative model-based FAS using databases from specialized and non-specialised databases.
2. To study the racial bias using an anomaly detection model on FAS databases and non-specialized databases.
3. To identify the most influential factors in racial bias.

1.5 Research Methodology

The proposed research will look into a variety of approaches that can be taken to address the problems that have been outlined above.

The initial stage of the research, described in Chapter 4, focused on enhancing the efficiency of a particular attack on face recognition systems and assessing its effect on FAS model accuracy. To achieve this, an image processing technique was employed to create a novel attack. The evaluation involved investigating the potential conversion of the developed adversarial attack into a direct presentation attack. ResNet50, a widely utilized deep neural network architecture in FAS, was chosen for this evaluation due to its capability to learn intricate feature representations from high-dimensional face images.

During the second stage of the research in Chapter 5, the primary challenge in developing a reliable FAS algorithm was dealing with the multitude of presentation attack types that the system needed to identify, along with other factors such as the quality of printed photos, screen resolution, and display type. All of these factors can significantly impact the performance of any FAS algorithm. As a result, a robust FAS algorithm must be capable of detecting previously unknown attack methods that were not anticipated during its deployment. In this context, we identified anomaly detection as an increasingly popular approach to FAS. This approach only trains the model on client images [27, 35]. The goal is to detect anomalous samples that do not fit the expected class distribution, thereby identifying potential presentation attacks. Thus, Convolutional Autoencoder (CAE) is used to evaluate the efficiency of the reconstruction of the training dataset using wild images and non-specialized databases. Anomaly scores are utilised during the testing phase to establish whether a sample belongs to the client class. This is achieved by comparing each sample's anomaly score to a predefined threshold. Client samples often have anomaly scores below the threshold, whereas attacker samples have scored over the threshold.

In the third phase, discussed in Chapter 6, we conducted an evaluation of various anomaly detection models using multiple FAS databases. We compared different feature extraction methods to identify effective feature representations. As a result of this investigation, we developed a model that utilizes a transformer and ResNet18 as

the backbone and a decoder for image reconstruction. To the best of our knowledge, this is the first study in the field of FAS to incorporate an anomaly detection model with a transformer architecture.

In the final chapter, Chapter 7, we explore the issue of racial bias in face spoofing detection. To conduct this study, we employ the Vector Quantized Variational Autoencoder (VQ-VAE) for anomaly detection in FAS. Then, we investigate the presence of racial bias using statistical and algorithmic tests. This study represents the first examination of racial bias in the context of FAS and is one of the most comprehensive investigations of bias in binary classifiers.

1.6 Contributions

In this thesis, the types of attacks we focus on are photos and videos, which are more common and easier to deploy than other attack modes. In all cases, we assume that the camera of the face recognition system acquires a single photo of the presentation session or records a short video from which we then extract a single frame and process it, aiming at detecting a possible presentation attack.

In Chapter 3, as discussed in Section 1.2, the traditional FAS algorithms have followed a binary classification approach to distinguish between genuine and fake faces. However, this approach necessitates a substantial amount of labelled training data and assumes that the attacker will employ a limited range of known attack methods. Consequently, this approach may prove inadequate when faced with more sophisticated and diverse spoofing techniques. From this survey, we offer a comprehensive understanding of the anomaly detection methods utilized in FAS.

In Chapter 4, we introduce a novel approach by creating imposter images that undergo specific image processing operations. Subsequently, we assess the performance of the Resnet50 algorithm against these attacks. Figure 1.3 illustrates examples of processed images that can successfully deceive FAS. Through these experiments, we aim to gain insights into the impact of processed images on the effectiveness of FAS algorithms. The key finding of our research is that the adversarial attack can be converted into a physical attack, highlighting the vulnerability of face recognition

systems to such manipulations.

Figure 1.3: From left to right: the first images are imposters from the database, and the processed imposter images with saturation values 63 and 223, respectively.

In the next stage, corresponding to Chapter 5 of the thesis, we evaluate the performance of an anti-spoofing method based on anomaly detection under the augmentation of the training set with in-the-wild images and images from non-specialised databases. In this chapter, we used an autoencoder as an anomaly detection method to reconstruct images which train with bona fide images only. We found that, under a cross-database testing protocol, the classifier's performance was higher with the augmented training set, indicating higher generalisation power. Figure 1.4 shows some examples from in-the-wild images that were used together with images from non-specialized databases to augment the training set. This approach can be particularly effective when there are many variations in the types of attacks, as the autoencoder can learn to recognize patterns in the data that are not associated with bona fide faces.

Figure 1.4: Some examples from the harvested wild images.

In Chapter 6, the proposed model AnoFormer combines the strength of the transformers and ResNet as backbones to extract features, which are then reconstructed by a decoder. The experiments show that the model's performance compares favourably to other one-class and two-class methods under certain evaluation protocols. Figure 1.5 illustrates the attention mechanism of the transformer model when processing an imposter image. The attention mechanism is a crucial component of the transformer architecture that enables the model to focus on relevant parts of the input image during the feature extraction process. Figure 1.5 illustrates the attention mechanism of the transformer model when processing an imposter image. The attention mechanism allows the transformer model to dynamically assign different weights to different regions of the imposter image based on their importance. The model can extract relevant representations for distinguishing between genuine and imposter faces by attending to specific facial features or regions.

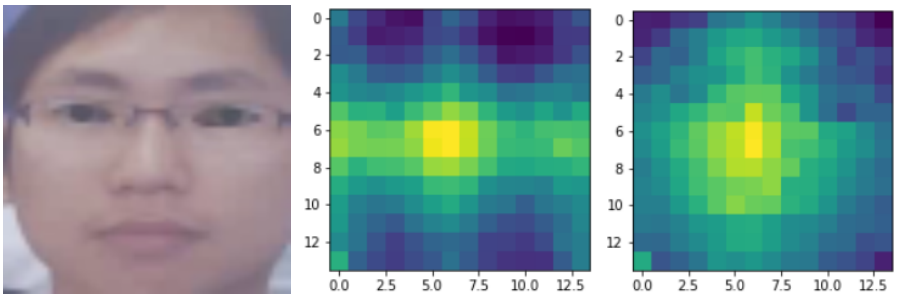

Figure 1.5: The imposter-cut-photo from Casia-FAS DB (left) and the visualization of its attention (right)

In the final section of our research, which corresponds to Chapter 7, we study the racial bias in a FAS model based on anomaly detection with a Vector Quantized Variational Autoencoder (VQ-VAE). We conduct a systematic, largely empirical, study of race bias based on a series of statistical and algorithmic tests. Our premises are that we are interested in the bona fide error; that binary outcomes, scalar responses, and latent space should all be analysed for bias; and that the threshold determining the classifier's operating point should be a variable rather than assumed to be fixed by a black-box procedure.

The main contributions of our work can be summarised as follows:

- A comprehensive review of anomaly methods for FAS.

- Testing the performance of one of the well-known algorithms proposed against processed images with colour manipulation.
- Evaluating anomaly detection method by reconstructing a training data set with in-the-wild images, a non-specialized database, and a face spoofing database.
- Proposing a novel model for FAS based on the anomaly method.
- A systematic analysis of racial bias in FAS.

1.7 Thesis Structure

The outline of the rest of the thesis is as follows: Chapter 2 delves into biometric traits, discussing various types of attacks against face recognition systems, common techniques for detecting such attacks, and the current state of the art in the presentation attack detection (PAD) problem. Chapter 3 presents a comprehensive survey of anomaly detection-based FAS techniques. Chapter 4 focuses on enhancing clients' images before using them for presentation attacks. The proposed technique involves colour manipulation to create new attacks by enhancing clients' images and degrading the performance of the Resnet50 model. In Chapter 5, including in-the-wild images and images from non-specialized databases in the training set is proposed, leading to a more robust auto-encoder. Chapter 6 introduces a new anomaly detection model for FAS called Anoformer. The model utilizes a transformer and ResNet18 as backbones and a decoder. In Chapter 7, investigating racial bias in FAS using VQ-VAE and statistical analysis is the main focus of this chapter. Chapter 8 concludes the thesis, summarises the findings, and offers several suggestions for further research.

CHAPTER 2

Background on Biometrics

Over the past few decades, spoofing attacks have become as a significant concern for biometric systems. Researchers have developed various techniques to combat these attacks and mitigate the risk of unauthorized access to biometric verification systems. It is essential to progress through several stages to understand biometric spoofing and develop practical anti-spoofing systems. The first stage involves defining spoofing attacks and examining their development and execution. Based on this understanding, suitable countermeasures can be formulated in the second stage. This chapter begins by providing general information about biometric traits in Section 2.1, followed by a discussion on face recognition in Section 2.2. Section 2.3 and Section 2.4 explore spoofing attacks, how they are created, and how they are executed. Using this knowledge, Section 2.5 describes the process of creating appropriate countermeasures to prevent spoofing attacks, providing an overview of the most effective anti-spoofing techniques for face recognition. Furthermore, Section 2.6 discusses the face anti-spoofing and face recognition databases used in this thesis. Lastly, Sections 2.7 and 2.8 present the evaluation metrics and protocols utilised in the field of face anti-spoofing.

2.1 Biometric Identification

2.1.1 Biometric Traits

Biometric traits are physical or behavioural characteristics that can be used for identification, authentication, or verification. Examples of physiological traits include fingerprints, facial features, iris or retina patterns, DNA, and voice, while behavioural traits include gait, typing rhythm and signature. Biometric traits are unique to each individual and are difficult to forge or steal, making them valuable tools for security and identification purposes.

Biometric authentication and verification technologies are increasingly used in various domains, such as border control, law enforcement, banking, and healthcare. A biometric system is a technology-based security system that uses physiological or behavioural characteristics to identify individuals. It can recognise and authenticate an individual's identity by analysing one or more unique physical or behavioural traits, such as fingerprints, face, gait, and signature. Compared to traditional identification methods, biometric systems have several advantages, including higher accuracy, greater security, and ease of use [13].

A biometric system comprises four essential parts: the sensor, the feature extraction module, the matcher module, and the system database module [36]. The primary functions of a biometric system are enrollment, verification, and identification. During enrolment, a user's biometric trait is recorded using a suitable sensor, such as a camera for face inputs, and then added to the system's database. Subsequently, feature extraction is used to extract salient characteristics from the data. These extracted features are combined with pre-defined identifiers, such as numbers or names, and are saved in the database as a template.

To complete the authentication process, the user submits a query to the system; this query consists of a unique sample shown to the sensor and compared to the previously saved template using the matcher, giving a match score that indicates how similar the template and query are. The system accepts the identification claim if the match score exceeds a preset threshold [36].

Biometric systems are also characterised as pattern recognition systems that

compare biometric data acquired from users after feature extraction with user feature data already saved in a database. Biometric systems can carry out two modes of operation: verification and identification.

Verification involves a one-to-one comparison between the individual's data and the data stored in the database. This comparison aims to prevent multiple users from accessing the same identity account. The verification task verifies the user's identification by matching the acquired biometric to the system's database. The identification mode, on the other hand, is used for negative recognition. In this mode, the database is checked for a match to the user's template, aiming to identify the individual based on their biometric data. In general, identification tasks can be summarised as follows: the user is recognised through a one-to-many comparison search of their template across all individuals in the database. When verification errors occur in biometric systems, they can be either False Matching Errors (FMRs), where an imposter is accepted, or False Non-Matching Errors (FNMRs), where a legitimate user is not accepted [37]. Biometrics are considered more trustworthy than other access mechanisms because, unlike physical access items such as cards, keys, and tokens, they are impossible to misplace, copy, steal, or forget at home. Passwords, PINs, passcodes, pass patterns, and other forms of non-physical access are all vulnerable to being lost, stolen, hacked, or shared and thus, applications such as security systems, consumer devices, and access control systems are increasingly adopting biometrics due to the security and convenience they offer.

2.2 Face Recognition Systems

Face recognition has emerged as a well-established field of study, with advanced techniques achieving recognition rates comparable to human performance under similar conditions. Face recognition [17] requires the least amount of personal information from the enrolled individual among the biometric methods. It is one of the best ways to identify and verify people. It is, therefore, one of the most popular biometric modalities and is included in various cutting-edge scenarios, especially in video surveillance. Three primary categories are used to classify facial recog-

nition methods: (1) local approaches,(2) holistic approaches, and (3) hybrid approaches [38]. Local approaches classify facial features without taking the entire face into account. Holistic approaches involve utilizing the whole facial area as the input for face recognition algorithms. Hybrid approaches employ local and global features to increase facial recognition accuracy. Extracting holistic facial features is usually done through methods classified into linear and non-linear. The most popular linear technique is Eigenfaces based on principal component analysis (PCA) [39]. In this approach [40], the Eigenfaces or face templates are generated from the principal components obtained through PCA. The PCA method reduces many variables that may be connected to a smaller number of variables (the principal components, PC). These PCs can be used as Eigenfaces or templates in face recognition.

On the other hand, in contrast to the linear approach of linear discriminant analysis (LDA), a non-linear method called kernel linear discriminant analysis (KDA) can be employed to enhance the performance of LDA. KDA is similar to the kernel extension of principal component analysis (PCA) and aims to capture non-linear relationships in the data. Arashloo et al. [41] proposed a non-linear binary class-specific kernel discriminant analysis classifier (CS-KDA) by drawing inspiration from spectral regression KDA. The CS-KDA classifier is designed to address the limitations of linear methods when faced with non-linearly separable data.

In local approaches, a structural classifier provided information regarding the location and shape of specific facial features, including the eyebrow curve, eyes, mouth, chin, nose, and lips [17].

The authors of [42] developed a system to aid facial identification even under challenging conditions, such as variations in expression and lighting. This system used local binary pattern (LBP) and K-nearest neighbours (K-NN) methods. LBP, being invariant to image rotation, has gained popularity as a preferred technique for facial identification.

Hybrid approaches combine the strengths of both local and holistic approaches in face recognition. These approaches leverage local feature extraction techniques to capture fine details while also incorporating subspace methods to address the high dimensionality of the feature space. In a study by [43], a mixed face recognition system

was proposed that integrates global and local features, namely Principal Component Analysis (PCA), Local Gabor Binary Pattern Histogram Sequence (LGBPHS), and Gabor wavelets. The PCA method is employed to reduce the dimensionality of the feature space. Subsequently, the local LGBPHS method is utilised for the identification stage. This design aims to simplify the Gabor filters and enhance the system's overall efficiency.

Deep learning has been the subject of extensive study to improve facial recognition. Lately, Convolutional Neural Networks (CNNs) have been applied to many classification issues to enhance accuracy [44]. The deep learning work on face recognition by Hu et al. [45] introduced a new method for evaluating the performance of the face verification method in the wild, which is a discriminative deep metric learning (DDML). DDML aimed to learn a Mahalanobis distance measure that would maximise inter-class variances while minimising intra-class variations. Their method had very competitive face-verification performance on common face databases.

In the study conducted [46], the authors utilized Linear Discriminant Analysis (LDA), a supervised technique that employs training samples to construct the projection matrix for feature extraction. In contrast, deep neural networks are versatile and can be trained for both supervised and unsupervised applications. The authors also incorporated the k-Nearest Neighbors (k-NN) algorithm and the Support Vector Machine (SVM) into their research. Their results conclusively indicated that the DNN method surpasses the performance of the LDA method.

2.3 Vulnerabilities of Face Recognition Systems

The expanding use of face recognition in security applications has led to a growing interest in identifying and addressing these systems' potential weaknesses and limitations. As shown in Figure 2.2, there are two categories of illegitimate access types to face recognition systems: direct and indirect attacks. The security integrity of such systems can be compromised at various attack points, as depicted in Figure 2.1. According to the source [1], eight potential attack points against biometric verification systems have been identified. Direct attacks are carried at the sensor

level, represented as point 1 in Figure 2.1, whereby attackers present fake facial artefacts, such as digital images or videos displayed on a screen or as printed photos or present masks, makeup, or surgically modified features. Indirect attacks can occur at these vulnerability points 2, 3, 4, 5, and 6, as shown in Figure2.1, and these attacks can manipulate the communication channel, feature extraction or matching, or templates.

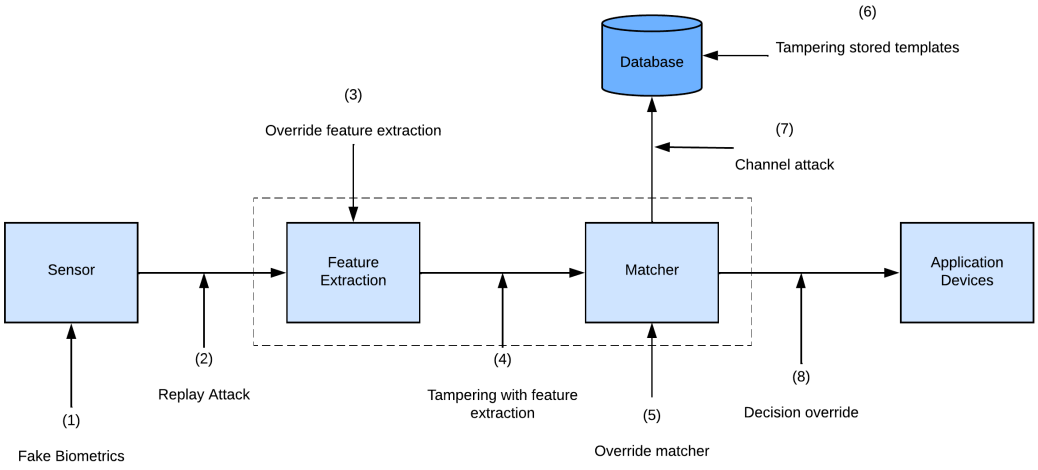

Figure 2.1: From 1 to 8, there are eight potential attack places. [1]

Since direct attacks require no pre-knowledge of the inner working of the system, attacks on the system using spoofed artefacts are becoming more common, especially with the impact of Covid-19, which has driven the widespread adoption of facial recognition systems [47]. Similar attacks are more accessible on a facial recognition system than other biometric systems due to the availability of facial information. Omar et al. in [21] tested the vulnerabilities of various robust applications, which were found to be easily bypassed by spoofed faces despite their high accuracy in recognizing faces.

One of the face recognition vulnerabilities is spoofing attacks. Many factors determine the quality of spoofing attacks. Initially, the original sample is utilised to create the attack, which can be an image of a mugshot obtained with the user's permission or an online photo. The quality of the input for facial recognition can also differ depending on the conditions of the spoofing attack, such as the illumination level. Spoofing attacks are carried out by the attacker, who controls the construction process, which includes selecting the spoofing medium, materials, tools, and

equipment required for the attack.

Each type of spoofing attack works differently and is characterized by unique properties. Therefore, developing effective countermeasures requires understanding the type of spoofing attack. Spoofing attacks can be either static or dynamic. Face identity is preserved in static spoofing attacks, but only a still image is shown. At the same time, dynamic spoofing attacks keep the appearance and liveliness of the human face by mimicking usual expressions and movements. Face spoofing attacks can be either two- or three-dimensional.

Face spoofing attack methods can involve a database of spoofed faces, which have been found to fool existing face verification systems. There are various ways in which the basic types of attacks can differ, and these differences may or may not be influenced by the attacker's intentions. For instance, the environment where the original sample was taken could be controlled or adverse. How the attacker holds the spoofing artefact can also be fixed or hand-supported [48]. In the latter, the attacker's involuntary hand movements can give the static attacks some life.

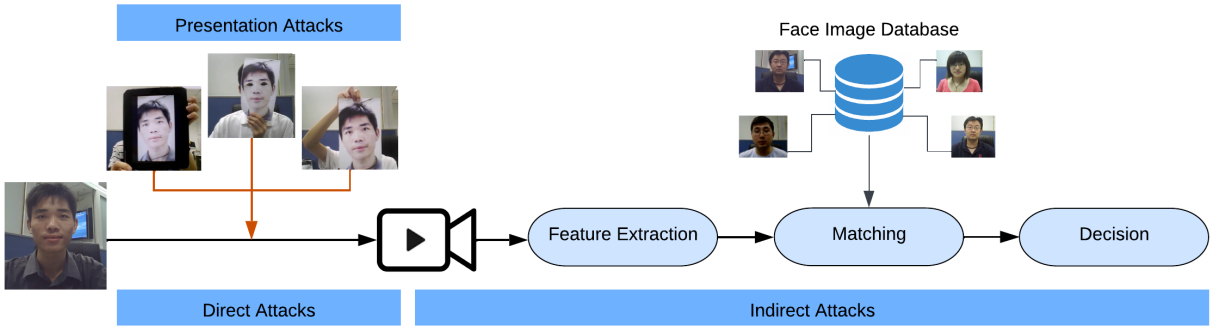

Figure 2.2: Illegitimate access types to face recognition systems: direct and indirect.

2.4 Types of Presentation Attacks

It is crucial to have a solid understanding of the most frequently used types of presentation attacks. Also, we need to know how they operate and what flaws in the system attackers are looking for specifically. Since photos of people's faces are easily accessible on social media, the most basic kind of attack in face spoofing uses

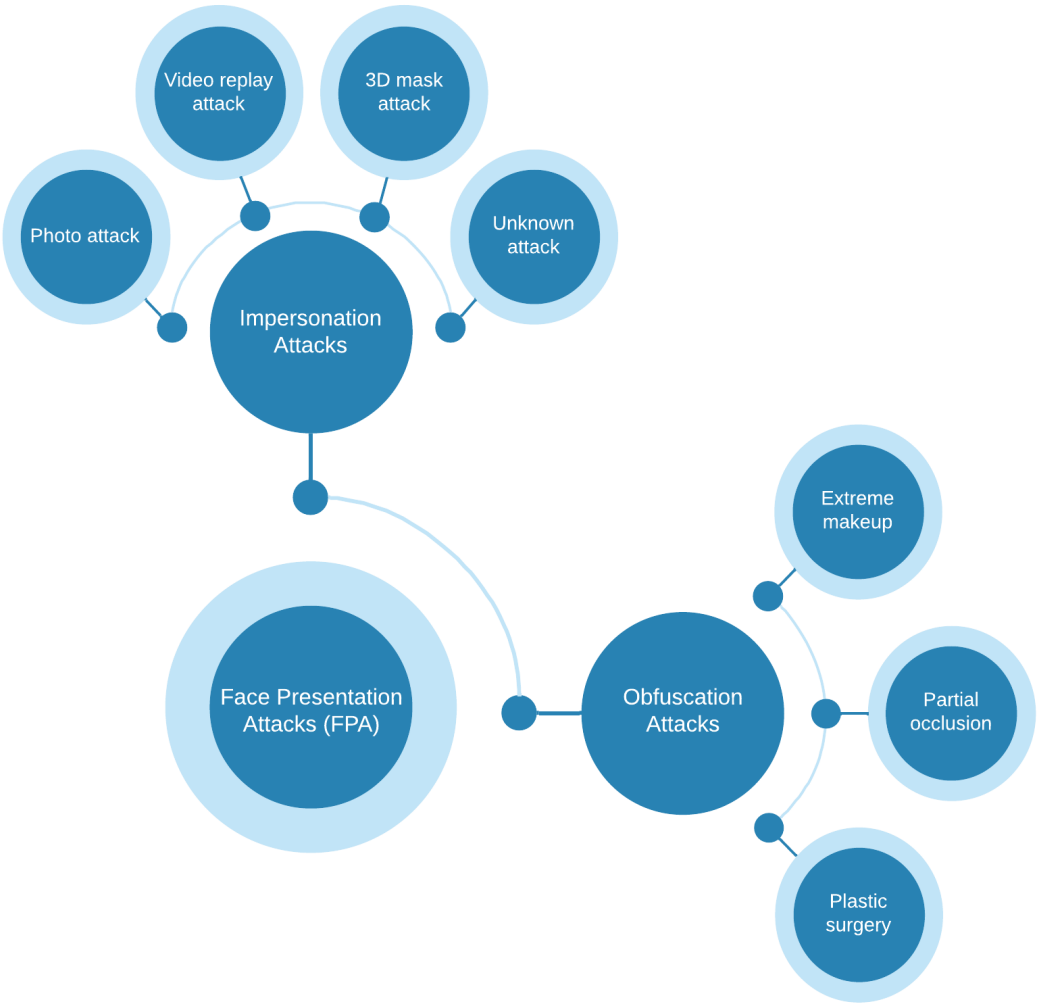

Figure 2.3: Types of face presentation attacks.

this approach.

These dynamic and static attacks have prompted the development of more sophisticated techniques. [49] divided spoofing attacks into two distinct categories, as illustrated in Figure 2.3. The initial classification is impersonation, where a person presents false representations to be identified as someone else by mimicking the physical features of a legitimate user through mediums like a photo, electronic screen, or 3D mask.

Photo attacks come in one of two types, namely eye-cut photo attacks and warp photo attacks. In the eye-cut photo attack, photos with holes in the eye region are printed to simulate blinking and eye movement. In contrast, the warped photo attack

involves printed photos bent in different directions to mimic face movement. A replay attack uses a looped video, giving it a more natural appearance than a still photo, and 3D mask attacks involve using three-dimensional facial masks to impersonate a genuine face. The second category involves obfuscation, which includes eliminating the attacker's identity through different means, such as wearing glasses, applying makeup, using a wig, and altering their appearance.

The primary distinction between these two categories of attacks is the attacker's objectives. In obfuscation attacks, the attacker endeavours to hide their genuine face, making it challenging for the face recognition system to recognize it as fake. However, in impersonation attacks, the attacker tries to present themselves as someone else in an attempt to gain access to resources or systems that are restricted to that person.

The second taxonomy for presentation attacks is based on dimensions, separating them into 2D and 3D varieties. Facial characteristics are presented to the sensor using a photo or video for 2D presentation attacks [50]. These attacks commonly use printed photos that lie flat or are wrapped, photos that crop out the victim's eyes or mouth, and digital replays of videos. With advances in 3D printing, a new, particularly effective form of presentation attack known as a face 3D mask [51] has emerged. Face masks are more realistic than regular 2D presentation attacks because of their richer colour, detailed textures, and precise geometrical shapes. Masks can be either hard or soft. Hard masks are commonly made from paper, resin, plaster, or plastic, while soft masks are usually made from materials like silicon or latex.

Most of the previous research in the field assumes that we have enough information on spoof artefacts and acceptable data samples to train models. However, the reality is much more challenging and complex. One primary concern is that attackers may employ novel attack techniques that have never been encountered before [52, 53]. Therefore, evaluating the effectiveness of facial Presentation Attack Detection (PAD) approaches against unknown attacks is essential.

2.5 Face Anti-spoofing Methods

Face anti-spoofing methods are techniques and algorithms to detect and prevent facial spoofing attacks. These methods rely on binary classification algorithms that differentiate between genuine faces captured by the camera, and imposter faces in printed photos, videos, or masks, which are used to deceive facial recognition systems. These approaches were created as countermeasures to spoofing attacks. According to the literature, the terms "face anti-spoofing," "face liveness detection," and "face presentation attack detection" are often used interchangeably. Nonetheless, the term liveness test may not always be suitable, such as when makeup is used to attack a face recognition system.

While FAS methods have made significant progress in recent years, it is evident that they cannot yet be regarded as fully developed technologies. The use of still imposter images, which is the most straightforward method to attack a face recognition system, remains highly effective at the consumer level.

Over the past few years, there has been a significant increase in the number of FAS methods. These methods can be divided into two categories: intrusive [54] and non-intrusive [55], depending on the extent to which they interfere with the process of acquiring biometric data.

Intrusive face anti-spoofing methods can have some disadvantages compared to non-intrusive methods. Firstly, they may require additional hardware or equipment, such as sensors or devices for measuring physiological responses, which can increase the cost and complexity of the system. Secondly, they may not be user-friendly and can cause discomfort or inconvenience to the user. For example, some methods may require the user to perform specific actions or movements or wear a specific accessory. As a result, non-intrusive approaches have gotten increased attention in the literature and are also the focus of our research.

Depending on the type of information utilised, current face anti-spoofing methods use two types of information: methods based on spatial information, which are called static approaches, and methods based on spatial-temporal information, which are called dynamic approaches.

This thesis examines two primary categories of face anti-spoofing methods: software-

and hardware-based. Software-based methods refer to techniques that focus on how the classification algorithm processes image features. [28], the authors classified face anti-spoofing techniques based on deep learning into three groups: deep learning, hybrid, and generalized methods. We divide the software-based methods the same as them. However, for a more comprehensive review, we add the traditional face anti-spoofing methods using hand-crafted features under this group.

On the other hand, hardware-based methods utilize additional hardware components or devices to enhance the face anti-spoofing capabilities. These methods leverage specialized sensors or technologies, such as infrared cameras, 3D depth sensors, or multi-modal systems, to capture more detailed and reliable biometric information for accurate spoofing detection. Figure 2.4 illustrates the taxonomy of our research in the field of Face Anti-Spoofing (FAS).

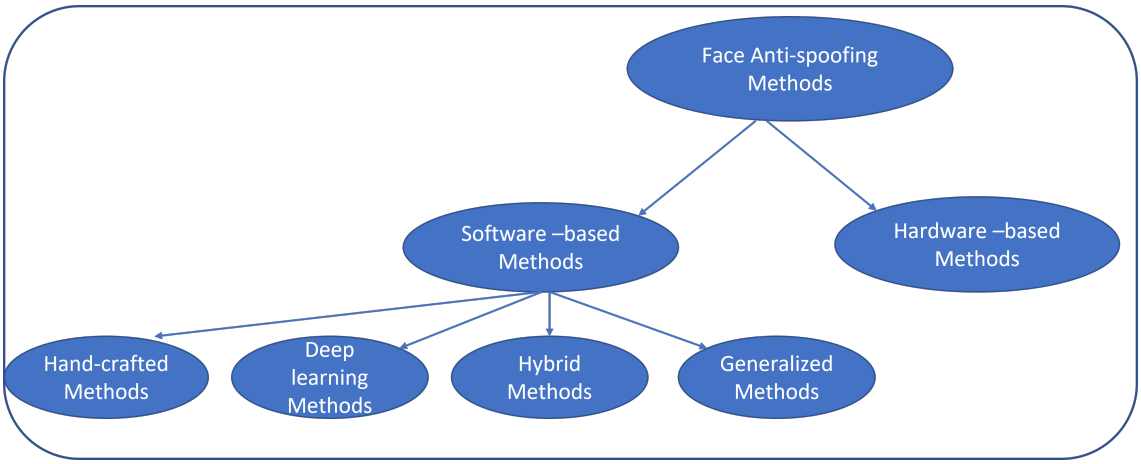

Figure 2.4: Our research taxonomy of face anti-spoofing methods

2.5.1 Software-based Methods

Handcrafted features-based methods

Face anti-spoofing methods using hand-crafted features are traditional methods that rely on feature engineering to detect spoof attacks.

Tan et al. [5] employed the Lambertian reflection model to investigate disparities between genuine human faces and fake ones. They based their analysis on two key observations: Firstly, real human faces exhibit three-dimensional characteris-

tics, unlike the two-dimensional representations of imposters. Secondly, discernible variations in surface texture exist between authentic faces and flat image surfaces. To implement the Lambertian reflection model effectively, the authors introduced two distinct techniques for extracting the underlying texture. The first approach involved a variational Retinex-based method, while the second utilized a Gaussian-based approach. To enhance both the speed and accuracy of classification, Tan and his colleagues made two crucial modifications to the sparse logistic regression model. These modifications included sparse low-rank bilinear logistic regression and the incorporation of a nonlinear model with empirical mapping.

The most commonly used feature for analyzing printed images in FAS is the Local Binary Patterns (LBPs). A prominent example is Maatta et al. [2], who used a multi-scale (LBP) approach to examine the texture features utilised for identifying flaws in the quality of printed faces. Figure 2.5 shows Maatta et al.'s proposed approach, outlining the various steps involved. Maatta et al. used three LBP configurations $\left(LBP^{u2}_{8,2}, LBP^{u2}_{16,2}, LBP^{u2}_{8,1}\right)$ and concatenate the histograms into a single feature vector. This vector was fed into an SVM classifier to distinguish between client and imposter images.

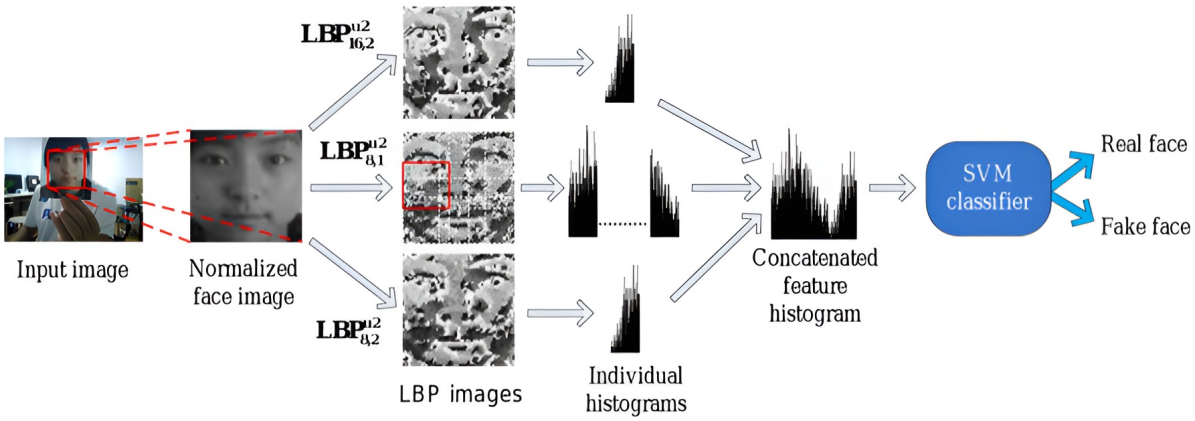

Figure 2.5: An illustration of the method proposed by Matta et al. [2]

In another work using LBPs, Chingovska et al. [6] employed a texture-based method to investigate the performance of LBPs in detecting face spoofing attacks. The authors evaluated different variations of LBP, including the original LBP, uniform LBP, and rotation-invariant LBP, on several publicly available datasets. The experimental results showed that LBP features effectively detect face spoofing at-

tacks, especially in combination with other feature extraction methods. However, the authors also highlighted the limitations of LBPs, such as their sensitivity to noise and illumination variations. Overall, the paper provides insights into the effectiveness of LBP in face anti-spoofing.

Boulkenafet et al. [3] proposed a FAS method based on colour texture analysis and concentrated on the luminance and chrominance features of the face image, as they are helpful for distinguishing the genuine from the fake faces. They used LBP to extract the features from individual image channels in different colour spaces, such as RGB, HSV, and YCbCr, as in Figure 2.6. These features were then tested on the CASIA and Replay-Attack databases to evaluate the effectiveness of the proposed approach.

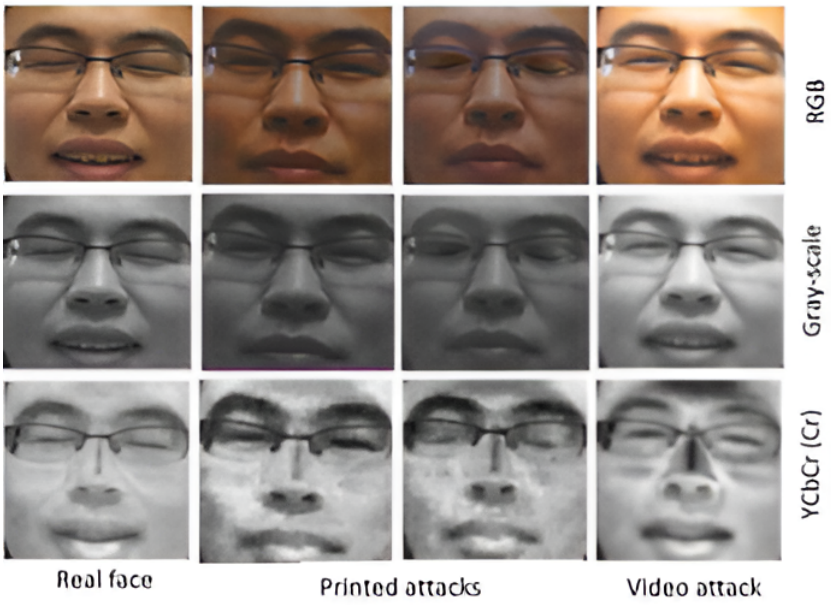

Figure 2.6: Some examples of genuine faces and print and video representations are presented in RGB, grayscale, and Y CbCr colour spaces, respectively [3].

Because handcrafted features are designed based on prior knowledge about specific attacks, datasets and scenarios, they often have limited representational ability, which can hinder the transferability of the model to new scenarios. Moreover, selecting appropriate handcrafted features can be time-consuming and require domain expertise. Furthermore, these features may not be optimal for discriminating between clients and all the different types of attacks, which limits their effectiveness in

real-world scenarios. Face anti-spoofing is a complex task that requires the accurate detection of subtle differences between genuine and fake faces. Traditional methods that use handcrafted features and machine learning algorithms can struggle with this task because they may not capture the full range of variations in the data. As we will see in the next section, deep learning is often more successful in face anti-spoofing because it can automatically learn and extract high-level features from a dataset, allowing it to capture subtle differences between genuine and fake faces that traditional methods may not detect.

Deep learning-based methods

FAS methods have evolved rapidly over the past decade from handcrafted methods to deep learning methods. According to recent research, deep learning methods are more effective than handcrafted features in FAS tasks. Meanwhile, the approaches are quite efficient within intra-dataset protocols but less useful in multi-domain datasets since they cannot easily adapt to new conditions [56].

When handcrafted features are applied in various environments or conditions, their limitations become more noticeable. This highlights that they may not work as effectively in diverse scenarios, where adaptability is a challenge.

These environments can include variations in illumination, camera angles, and backgrounds that can affect the quality of the captured data. Moreover, high-level (deep) features must be extracted from a dataset using multi-layered methods to handle complex tasks.

One of the first attempts at FAS with CNNs was by Yang et al. [57]. Their work consisted of two phases; the first utilised a CNN to extract features from the input images, and the second utilised SVM to classify the features as real or fake. Also, they used different data augmentation strategies. Their model performed well in the intra-testing stage, where the suggested method obtained HTERs (half total error rates) less than 5% on two datasets. However, it should be noted that the proposed method still cannot achieve satisfactory performance in the inter-test protocol. As discussed earlier, due to varying conditions, variation in performance between different datasets is inevitable.

A new approach was introduced by Wen et al. [58] based on image distortion analysis, whereby they analysed the distortion patterns in the image to distinguish between real and fake faces. The authors showed four types of IDA features (specular reflection, blurriness, colour moments, and colour diversity) used to catch distorted images in spoofed face photos. They used an ensemble SVM classifier to classify the features as real or fake. Combining these four features resulted in an IDA feature vector with a dimensionality of 121.

Atom et al. [4] proposed an auxiliary depth map to improve the results of face anti-spoofing. Their fully convolutional network (FCN) was trained using depth maps to extract local and global features of images as the input of two networks. Here, the local features are extracted from random patches within the face region, while the depth features leverage the whole face. The authors argued that 2D spoofing examples, such as printed papers and screens, lack 3D information, so they trained a CNN-based depth estimator that produces 3D depth maps for genuine faces and flat maps for spoofed faces. While the performance is enhanced with depth as auxiliary supervision, depth-based techniques might not be effective when a paper mask attack, which includes depth information, is carried out. Because it can have a similar depth map as a genuine face, making it difficult for depth-based methods to differentiate between them.

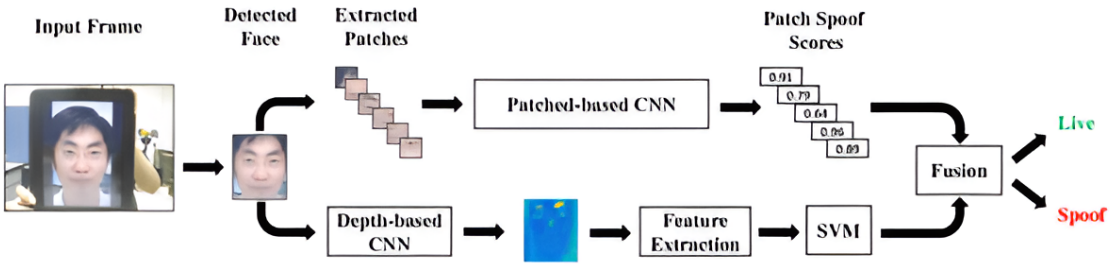

Figure 2.7: The proposed method by Atom et al [4].

From the transfer learning perspective, [59] conducted their experiments using various pre-trained models (Inception and ResNet) to detect spoofed faces. Different factors were considered, such as the model depth, learning rates and initialization of the weights. The authors evaluated the Inception-v3, ResNet50, and ResNet152

models under various conditions. The results indicated that ResNet152 is the best model for FAS when using only the dense layers trained with ImageNet weight transfer at 10^{-3} learning rate. ResNet152 improves with a lower learning rate, while ResNet50 does better with a higher learning rate. According to the findings of these tests, transfer learning with deep models produces superior results to using these networks with random weights or beginning the learning process from scratch.

Using Spatio-Temporal Anti-Spoofing Networks (STASN), Yang et al. [60] demonstrated the ability to detect presentation attacks in photos and videos. STASN is composed of three modules, namely Temporal Anti-Spoofing Module (TASM), Region Attention Module (RAM), and Spatial Anti-Spoofing Module (SASM), representing temporal, regional, and spatial anti-spoofing, respectively. The suggested TASM uses CNN and LSTM units to learn the temporal properties of the input video. The experiment outcomes using public FAS datasets are competitive with the state-of-the-art.

Zhang et al. [61] proposed a Face PAD model to disentangle features using depth and translated images. They solve the problem of FAS by using disentangled representation learning. This method divides the latent representation into liveness characteristics and content features. Together, the LBP map and depth map regularise the liveness feature space, which is essential in identifying genuine people from spoof patterns.

Using the Central Difference Convolutional Networks (CDCN), the authors in [62] argued that this operator could capture more fine-grained details and spatial information than traditional convolutional networks. The CDCN architecture comprises multiple CDCN blocks, each containing two main components: a central difference convolutional layer and a gating mechanism. The central difference convolutional layer applies the central difference operator to the input feature maps to extract more detailed information. At the same time, the gating mechanism helps to suppress noise and highlight important features. The authors also proposed a variant of CDCN, which combines the CDCN blocks with the self-attention mechanism. This allows the model to capture better the input data's long-range dependencies and contextual information.

Most of the studies presented earlier imply that there are sufficient data samples and information regarding spoofs to train a model. The significant progress in deep learning has made it possible for most cutting-edge face PAD algorithms to display state-of-the-art results in intra-database tests on the many public datasets. It is still difficult to generalise to scenarios involving several datasets, mainly because the test set may contain face PAs that were not present in the training data, which presents a challenge when attempting to generalise.

Hybrid Methods

Hybrid methods combine handcrafted and deep learning features to improve spoofing detection performance. In the study conducted by Asim et al. [63], they combined CNN and handcrafted LBP-TOP technique were employed to extract features. The classifiers were then trained using spatio-temporal information from video sequences to differentiate between genuine access and various spoofing attacks. Comparing their results to state-of-the-art, they achieved competitive results.

Another novel hybrid technique proposed in [64] which three discriminative representations; firstly, they used the Spatial Pyramid Coding Micro-Texture (SPMT), which can be capable of extracting the local appearance information, then utilised a deep learning framework, namely Single Shot MultiBox Detector(SSD) for detection and extract context cues from the face images. Finally, they employed the designed descriptor Template Face-Matched Binocular Depth (TFBD) feature to characterise the stereo structures of client and imposter faces. This descriptor considered the face's depth information and geometric characteristics, providing additional cues for accurate FAS. The authors proposed two different Combinations of representations. They first introduced a decision-level cascade technique that combined SPMT with SSD, and they employed a basic score fusion strategy to fuse the face structure cues (TFBD) with the local micro-texture features (SPMT). They demonstrated their effectiveness through extensive evaluations of multiple datasets.

Additionally, in [65], a perturbation layer was introduced as a trainable pre-processing layer for low-level deep features. The point of this layer was to improve the discriminative capability of the deep features by introducing perturbations or

modifications to the feature representations. The described layer utilised the deep features from a candidate layer in a CNN and the corresponding handcrafted features from an input image. It generated adaptive convolutional weights that assign significance to the pixels in the deep features of the candidate layer.

The authors in [66] suggested using combined local binary pattern (LBP) and simplified Weber local descriptor (SWLD) encoded CNN models to detect face-spoofing. LBP captured the orientations of edges but disregarded intensity information, while SWLD preserved local intensity information but omitted edge orientations. The authors employed a non-linear SVM classifier with a kernel function to determine whether the input is a client or a spoofed face. The approach achieved promising results in terms of performance on benchmark datasets.

Hybrid methods are effective in face anti-spoofing. However, these methods have some limitations, such as overfitting, which results in the poor generalization of new data. Also, the complexity can be increased by combining the handcrafted features and deep learning because the model architecture would be more complex, leading to longer training times, increased computational requirements, and difficulty interpreting the model. These limitations and all the previously mentioned handcrafted feature limitations in section 2.5.1 make this method impractical.

Generalized Face Anti-Spoofing Methods

Conventional FAS approaches based on deep learning may not generalise well to unseen environmental or sample characteristics, e.g., lighting and camera resolution and unknown attack types. As a result, these technologies may be unsuitable for scenarios that require extreme safety.

As a result, more researchers are focusing on improving the generalisation capability of deep FAS models. On the one hand, domain adaptation is utilised to achieve robustness in FAS of limitless domain variations. On the other hand, anomaly detection frameworks and zero-shot or few-shot learning frameworks are used to detect unknown face presentation attack types.

In the domain adaptation aspect, when deep models are trained directly on various source datasets, such as WMCA [67] or CASIA-MFSD [68], significant domain

changes between the source and target domains can easily lead to poor performance on a biassed target dataset, such as SiW [7]. The domain adaptation strategy utilises the knowledge from the target domain to overcome differences between the source and target domains.

The authors in [69] suggested a unified unsupervised and semi-supervised domain adaptation network (USDAN) for FAS to minimise the distribution difference between the domains. The model combines two learning approaches: unsupervised and semi-supervised domain adaptation, which are marginal distribution alignment module (MDA) for domain-invariant feature learning and conditional distribution alignment module (CDA)for centroid alignment of labelled features. The proposed model comprises three essential parts: a feature extractor, a domain adaptation layer, and a classifier. Their feature extractor (ResNet18) extracts meaningful features from face images. The results of the evaluation in the USDAN model were outstanding.

However, in many practical FAS scenarios, it is challenging and costly to collect a lot of data for training, particularly from spoofing attacks, which are required by the domain adaptation techniques to minimise the distribution discrepancy between the source and the target domain. Generally, a FAS model can be fine-tuned with only samples of new attacks and become able to detect novel attacks. Nevertheless, as spoofing techniques continue to develop, collecting even a small number of data for each new attack can become time-consuming and resource-intensive.

Liu et al. [70] proposed a Deep Tree Network (DTN) to address the challenge of unknown spoof attacks. The DTN was created to categorise the spoof samples based on their semantic similarities to known attacks. By analysing the similarity of the embedding features, the DTN dynamically routed known or unknown Presentation Attacks (PAs) to the corresponding spoof clusters. However, the DTN's performance in identifying and classifying such unknown attacks was limited because the absence of prior knowledge about these novel attacks hindered the network's ability to effectively distinguish them from genuine samples, reducing its overall accuracy and reliability in face anti-spoofing.

Qin et al. [71] proposed a novel method called Adaptive Inner-update Meta

Face Anti-Spoofing (AIM-FAS). AIM-FAS used meta-learning techniques to train a meta-learner, particularly one focused on detecting unseen spoofing types. The meta-learner learns from a combination of predefined living and spoofing faces and a few examples of new attacks. This approach enables the detector to handle zero-shot and few-shot learning scenarios in face anti-spoofing. Experimental results demonstrate that AIM-FAS outperforms existing methods and performs better in existing zero-shot face anti-spoofing protocols. However, applying few-shot meta-learning to new attacks may result in catastrophic forgetting of prior attacks.
In both zero-shot and few-shot scenarios, the models demonstrated increased robustness in effectively managing predefined Presentation Attacks (PAs) as well as a limited number of new, previously unseen PAs. This approach aimed to improve the model's ability to generalise to unseen attacks by leveraging a combination of existing knowledge and limited further information. Despite the few-shot and zero-shot learning that has shown promising results for detecting unknown attacks, the performance of the models decreases when the data of the target attack types cannot be adopted.

Despite the large amount of research on face anti-spoofing for face recognition systems. To the best of our knowledge, there is no generally accepted PAD method capable of defending against all kinds of attacks. However, working with anomaly detection methods can provide several advantages. One of these is improving the generalisation of the FAS model because it would learn only clients' images and not be limited to specific types of attacks. It can identify various spoofing techniques, including print, replay, and 3D mask attacks. Additionally, anomaly detection can be customised to fit different scenarios, making it a flexible technique for face anti-spoofing. Overall, anomaly detection with face anti-spoofing can provide a more robust and accurate face recognition system. Chapter 3 has a detailed review of existing work based on anomaly methods in the field.

2.5.2 Face Anti-spoofing Methods based on Special Hardware

Regarding applications requiring daily face recognition, FAS, which utilizes a commercial RGB camera, is a great solution that effectively balances hardware cost and security level, e.g., mobile login. Nevertheless, certain high-security circumstances, such as face payment, call for an extremely low rate of false acceptance errors. Recent years have seen the development of cutting-edge sensors that can perform various functions to facilitate an extremely secure FAS. The use of stereo cameras (VIS-Stereo) is advantageous over monocular visible RGB cameras (VIS) because they provide 3D geometry information for detecting 2D spoofing attacks [72].

Depth sensors such as Time of Flight (TOF) [73] and 3D Structured Light (SL) [74] have been built into most smartphones and provide accurate 3D depth distribution for detecting 2D spoofing attacks. TOF is more robust to environmental conditions than SL. near-infrared (NIR) imaging (900 to 1800 nm) [75] in face anti-spoofing, which is complementary to visible light and effectively exploits reflection differences between live and spoofed faces. However, it should be noted that NIR imaging has poor imaging quality at long distances.

The integration of visible (VIS) and NIR imaging hardware modules (VIS-NIR) is another high-performance and cost-effective solution for many access control systems. However, it should be noted that other niche sensors such as Shortwave Infrared (SWIR) [76] and thermal cameras [77] may be more effective for detecting generic spoofing attacks but may not perform as well under certain conditions such as when subjects are wearing transparent masks.

In the end, this group of approaches is frequently regarded as inappropriate due to the expense of the additional hardware. They might also be less practical from a deployment perspective because providing additional hardware for particular applications, like mobile devices, can be challenging.

2.6 Databases

2.6.1 Face Anti-spoofing Databases

In recent years, several databases have been developed to facilitate the research and evaluation of face anti-spoofing methods. In this context, here we cover in more detail the face anti-spoofing databases, especially used in anomaly detection work, such as the Replay-Attack [6], NUAA [5], and SiW [7] databases because this thesis focuses on the anomaly detection concept. Indeed, these databases have been widely used in the literature to evaluate the effectiveness of face anti-spoofing methods and have helped to advance the state-of-the-art in the field. A thorough understanding of these databases can aid with the development and evaluation of more robust and effective face anti-spoofing methods.

Table 2.1 summarizes these databases regarding subject numbers, modality, expressions, attack types, and number of subjects. By examining the characteristics of these databases, researchers can better understand the range of potential challenges and variations in the data that need to be addressed when training and developing FAS models.

Typically, a database is divided into three sub-groups: training data, development data, and test data. The training data is the data used to train the model. The model learns from this data and adjusts its parameters to minimize the training error.

The development data is a separate set of data that is used to evaluate the performance of the model during training and to make decisions about the model architecture, hyperparameters, and optimization algorithms; that is, the development data is used to tune the model and prevent overfitting the training data. This development set corresponds to what is commonly referred to as the validation set in the machine learning literature. The test data is the set of data used to evaluate the final performance of the model after it has been trained and tuned. The test data should represent the problem at hand but be completely independent of the training and development data to avoid overfitting.

Database	Year	Subjects	PA-types	Ethnicity	Pose	Expression	PAI	Modality
NUAA [5]	2010	15	P Wrapped	Asian	Frontal	No	A4 paper	RGB
CASIA-FASD [68]	2012	50	P Wrapped Cut R	Asian	Frontal	No	C iPad	RGB
REPLAY-ATTACK [6]	2012	50	P P-D R	Caucasian African Asian	Frontal	No	A4 paper iPad 1 iPhone 3GS	RGB
MSU-MFSD [58]	2015	55	P R	Caucasian African Asian	No	No	A3 paper iPad Air iPhone 5s	RGB
REPLAY-Mobile [78]	2016	40	P, R	No	Frontal	No	flat, monitor	RGB
Oulu-NPU [79]	2017	55	P R	Caucasian Asian	Frontal	No	flat,phone	RGB
SIW [7]	2018	165	P R	Asian Indian Caucasian African	[-90,90]	Yes	Samsung iPhone 7 iPad PC	RGB
Rose-Youtu [80]	2018	20	2D,3D	Asian	Frontal	No	A4, Mac screen Lenovo LCD screen	RGB
WMCA [67]	2018	72	2D/3D P R M	Asian Caucasian African Hispanic	Frontal	No	Paper mask, rigid mask, flexible/silicon mask, glasses, fake head	RGB/ Depth/ IR/thermal

Table 2.1: A summary of publicly available datasets was used in anomaly detection methods for face anti-spoofing. This table represents the year, the number of subjects, PA types, ethnicity, pose, expression and PAI modality in these DBs. In the column 'PA type', 'P, 'R', and 'M' denote 'photo' and 'replay-video' and 'masks', respectively.

NUAA

In 2008 Tan et al. [5] introduced the first public database in face anti-spoofing called NUAA. This database contains printed photo attacks involving 15 subjects and 12,614 samples which are 5,105 client accesses and 7,509 attacks. The genuine face images are captured using a generic webcam in three different sessions and lighting conditions. The resolution of these images is 640x480. Subjects were asked to keep a neutral facial expression and frontal pose and avoid blinking their eyes. All identity images in the NUAA database are of Asian people.

Figure 2.8: Imposters examples from NUAA DB [5].

There are two kinds of attacks in NUAA: print and warped print attacks, which are made in two different sizes on regular paper and photographic paper. Each client has several attack samples with various positions and distances from the recording equipment. Figure 2.8 shows examples of samples from this database.

CASIA-FASD

To increase the generalisation performance of trained models and more accurately reflect real-world scenarios, the authors of this research acknowledged the need for face anti-spoofing datasets that cover a wider range of potential attack variations and include a greater diversity of face types. Hence, this was the first publicly available face PAD dataset to include printed photos and video attacks [68]. This database contains 50 subjects at different resolutions and lighting conditions. Three types of spoofing attacks are covered by the CASIA-FASD database, namely replay, warp print, and cut print, and there are three different types of qualities: low, normal,

and high.

Replay-Attack

In 2012, Chingovska et al. [6] released the Replay-Attack DB, developed to address some of the shortcomings of previous spoofing databases. It was designed to be a more comprehensive and diverse dataset for research on spoofing attacks. The Replay-Attack DB has 1,300 videos of 50 participants' photo and video attacks under various lighting situations. This database consists of 300 client-access videos and 1000 videos of imposters holding photos or tablets playing video recordings. In any of the training, development and test datasets, a subject that appears in one of them does not appear in the others under any lighting conditions. The attacks have been divided into three categories: images printed on A4 paper, video and photo presented on an iPhone mobile and photos and videos presented on an iPad screen. The client access videos and the imposter ones were captured in two lighting conditions. As shown in 2.9, the uncontrolled setting depicts the office without the lights on and a complex background. On the other hand, the controlled setting portrays the office with the lights on, shades up, and a uniform background. The Replay-Attack DB comprises individuals from three racial categories - Caucasian, Asian, and African.

Figure 2.9: Replay-Attack samples from a controlled setting (top row, left to right) and an uncontrolled environment (bottom row, left to right) comprise clients, printed photos, mobile phone and tablet photo attacks [6].

MSU Mobile Face Spoofing Database

For this database [58], actual access and attack videos were used. This database includes spoofing attempts that were limited to mobile devices. It has 55 clients and 280 video samples, with 210 attacks and 70 real accesses. The videos were recorded under challenging lighting conditions. There are three primary sorts of attacks: print attacks, videos with two different qualities, and repeated videos for smartphones and tablets. The protocol provided by MSU-MFSD does not include separate enrolment information and merely consists of the train and test sets with non-overlapping identities.

Replay-Mobile

The Replay-Mobile Database [78] for Face Spoofing contains 1190 video clips of image and video attack attempts made for 40 distinct clients, each shot in a unique lighting environment. The videos in the Replay-Mobile Database were recorded using modern technology devices. Specifically, an iPad Mini2 running iOS and an LG G4 smartphone running the Android operating system.

OULU-NPU

OULU-NPU [79] was launched in 2017 to advance research in face anti-spoofing and improve the performance of anti-spoofing systems. This database has 55 subjects represented, with the images captured by the front cameras of six various smartphones. The photos in most earlier datasets were captured under restricted circumstances. However, this database has a wide range of motion, blur, lighting, backgrounds, and 35 different head postures. The photos were taken in three different settings (environment, facial artefacts, and acquisition devices), each with a unique lighting and background.

Spoofing in the Wild Database

Spoofing in the Wild Database (SIW) [7], released by Liu et al., includes high and low-resolution printed photos and videos created with four different presentation

attack instruments. There are 165 participants in SiW, each with eight live videos and up to 20 spoof videos. This results in 4,478 videos, all of which have 1080P HD resolution. As shown in Figure 2.10 that the live videos have distance, pose, illumination, and expression variations and were collected in four sessions. During the first session, the individual moved their head towards and away from the camera at various distances. In the second session, they altered the yaw angle of their head within $[-90°, 90°]$ and displayed diverse facial expressions. Sessions 3 and 4 consisted of repeating the first two sessions, but the collector moved the point light source in various directions around the individual's face.

Figure 2.10: Examples of live and spoof videos in the SiW DB top row and bottom row, respectively [7].

The SiW dataset includes two categories of face spoofing attacks: picture attacks and video replay attacks. Picture attacks consist of two types of print attacks, and video replay attacks have four types of attacks that utilize four spoof mediums (PAIs): two smartphones, a tablet, and a laptop. Various print sizes, poses, expressions, and lighting conditions were used with different devices to generate different qualities of print attacks. They also extract a frontal-view frame from a live video for lower-quality print attacks.

A further advantage of SiW over other publicly available anti-spoofing datasets is its racial variety. It is the only dataset that includes Indian participants and has the highest percentage of African Americans. The dataset encompasses males and females, and its samples are more representative of real life in terms of expression (including genuine emotions, happiness and surprise), position, illumination, and

resolution. Four different presentation attack instruments allow SiW to provide various presentation attacks, from high-quality printed graphics and movies to low-quality options. Print and replay attacks can utilise these media files, images, and videos can be utilised for print and replay attacks, respectively.

ROSE-YOUTU

ROSE-YOUTU [81] was published in 2018 and has 20 subjects. This database contains many lighting scenarios, camera makes, and attack types. It considers three different spoofing attack types: printed paper attacks, video replay attacks, and masking attacks. A4-sized facial images with still printed paper and quivering printed paper are utilised in the printed paper attacks. A face video is played on Mac and Lenovo LCD screens to simulate a video replay assault. Masks with and without cropping are taken into consideration for masking attacks.

Wide Multi-Channel presentation Attack

The WMCA was developed by George et al. [67] and consists of both 2D and 3D attacks from 72 different identities. The information in the database was gathered from a diverse selection of imaging sensors and accounts for varying degrees of illumination, recording settings, and other aspects. Colour, depth, infrared, and thermal channels record the data. The presentation attack types in this database may be categorised into seven groups: paper mask, rigid mask, flexible/silicon mask, glasses, fake head, print, and replay. The improved detection of presentation attacks can be accomplished by utilising features gained from multimodal images.

Findings on Face Spoofing Databases

Constructing spoofing attacks can be time-consuming and may need significant resources, while the type of attack produced requires particular manufacturing expertise. Furthermore, creating a comprehensive dataset covering various practical settings for face anti-spoofing in real-world scenarios is still challenging. Hence, collecting attack data for numerous clients may be a challenging and costly task, especially for specific types of attacks. [82].

As a result of difficulties in data collection, the current face PAD datasets exhibit constrained diversity in terms of presentation attack types, PAI, and acquisition instruments for genuine faces. This limitation in the performances of the existing face PAD methods can be partly attributed to the need for more diversity in the datasets, which limits the effectiveness of all hand-crafted features and deep learning-based approaches.

This research will employ an anomaly detection technique known as a one-class classification to overcome these limitations. Specifically, the focus will be on using only bona fide images during training, while the imposter images will be reserved for the testing phase. This approach is intended to enhance the generalisation of the anomaly detection system and improve its ability to identify presentation attacks.

Also, the data has clear data asymmetry, as only the client data are cheap and abundant. This thesis investigates whether this asymmetry is reflected in training and testing or whether the training and testing data pool is restricted to limited specialised databases.

Table 2.1 illustrates that PA-types refer to the types of presentation attacks included in the database, and PAI refers to the presentation attack instruments used. The pose is the orientation of the face in the images, and the modality relates to the type of image used, such as RGB, depth, or infrared.

The databases we choose must have various challenging conditions, such as variations in pose, illumination, and expression, racial representation and several different types of PAIs. This is why we chose these databases NUAA, Replay-Attack, and SIW as in Table 2.1 the grey raws. These are famous face anti-spoofing databases used in research for developing and evaluating face anti-spoofing methods. Overall, these databases are valuable resources and contain various challenging conditions and realistic PAIs that can help researchers improve the performance of their methods.

2.6.2 Face Recognition Databases

Previous research on anti-spoofing has primarily relied on face-spoofing databases. However, this thesis proposes using non-specialised databases from other fields, such

as face recognition, to combine diverse images. This approach will enhance the model's performance by utilising more diverse datasets. This thesis employs the face recognition databases Casia-WebFace [83], SCface [84], and Racial Faces in the Wild [85]. We later outline the rationale for utilising them in this thesis.

Casia-WebFace

The Casia-WebFace database [83] was chosen for this research due to its large size and diverse range of subjects. It is a collection of face photos maintained by the Institute of Automation of the Chinese Academy of Sciences for use in facial recognition research. It contains over 500,000 images of more than 10,000 subjects. Because the photos were obtained from the internet, they have a wide variety of lighting, position, and expression differences, as well as other elements that can affect recognition accuracy. Because of its diversity, the Casia-WebFace database is considered a challenging dataset. As a result, it is frequently used to evaluate the accuracy of new face recognition systems. It is also suitable for our purpose of enhancing face anti-spoofing methods as it incorporates diverse and non-specialised images.

SCFace database

The SCface database [84] was chosen for this research due to its more realistic representation of unconstrained face recognition scenarios. This database allows for a more comprehensive evaluation of the proposed method and its potential for real-world applications. The database contains facial photographs and has 4160 static images of 130 identities taken under various lighting and expression settings. The images were captured by multiple cameras of varying quality to simulate real-world conditions, with an emphasis on various law enforcement and surveillance use case scenarios. Researchers can test and evaluate facial recognition algorithms in different settings thanks to the database's use of varied illumination and expression circumstances. As in Figure 2.11 Some examples of this database which used in this research.

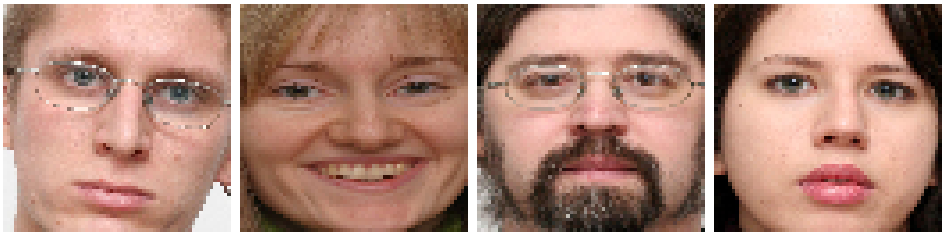

Figure 2.11: Examples from SCface DB.

Racial Faces in-the-Wild

A facial recognition database called Racial Faces in the Wild (RFW) [85] has been used to study prejudice against different races. Approximately 3000 people represent each of the four racial groups used in this study: Caucasian, Asian, Indian, and African as shown in Figure 2.12. They evaluate the algorithm's ability to distinguish evenly between races. Instead of downloading from websites, they randomly collected these images from MS-Celeb-1M [86]. Thus, they are suitable for measuring racial bias objectively. In contrast, when images are downloaded from websites, the selection process is often biased towards specific demographics or groups, which can introduce prejudice into the dataset. One potential use for them is to impartially assess and compare how well an algorithm recognizes individuals from various racial backgrounds.

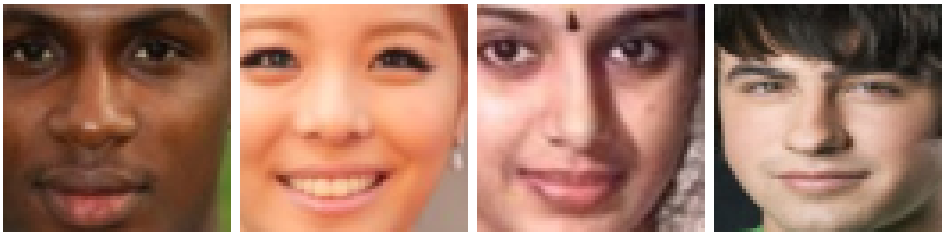

Figure 2.12: Examples from each race group in RFW DB.

2.6.3 Face Detection

Face spoofing DB Backgrounds

There are several reasons why most face anti-spoofing works do not include background information in their images. One reason is limitations in data availability: in some cases, the data collected for face anti-spoofing may not have background

information, either because it was not captured or because it was not deemed necessary for the task. Another reason is that some face anti-spoofing methods focus primarily on an individual's facial features and thus do not consider the background information. This is because facial features often provide enough information to distinguish between real and fake faces. Additionally, incorporating background information into a face anti-spoofing algorithm can increase the computational complexity of the system, slowing down the processing time and making the system more resource-intensive.

Face detection Methods

The Haar feature-based cascade classifier is a machine learning-based object detection algorithm that utilizes Haar-like features [87] to detect objects within an image or video stream. These Haar-like characteristics are simple rectangular features computed from the intensity values of an image. They capture information about contrast variations and edge patterns within the image. The Viola-Jones method [87] applies these features by sliding a window across the image and calculating them for each window. These characteristics are then used to determine whether or not the window contains an object.

The second method is MTCNN [88], or Multi-Task Cascaded Convolutional Networks, a deep learning algorithm designed to detect faces in images and videos. The system consists of three stages: the first stage detects possible face regions and eliminates non-face regions, the second stage refines the face region to more accurately locate the face, and the final stage outputs the faces landmarks, such as the nose, mouth, eyes, and the bounding box around the face. The MTCNN is known for its fast processing speed and high accuracy in face detection, making it a popular choice for various face-related tasks, such as facial recognition and face alignment. It can handle faces in multiple poses and lighting conditions, making it suitable for real-world applications.

In most of the work in this thesis, facial images are cropped and aligned before training using MTCNN to overcome obstacles, such as backgrounds that can cause a significant performance drop for face anti-spoofing, and then the images are resized.

2.7 Evaluation Metrics

FAS can be seen as a binary classification problem, and the performance is evaluated through various performance metrics. These FAS algorithms have two input classes, commonly called positive and negative classification models. They are evaluated according to the types of errors they commit and the methods they use to measure them. By definition, binary classification systems exhibit errors known as false positives and false negatives. Thus error rates are traditionally measured as the False Positive Rate (FPR) and the False Negative Rate (FNR). The FPR is the ratio of false positives to the total number of negative samples. At the same time, the FNR is the ratio of false negatives to the total number of positive samples. The terms FPR and FNR can also be replaced by the terms False Acceptance Rate (FAR) and False Rejection Rate (FRR) because the performance relies upon the acceptance or rejection of the sample in anti-spoofing systems. The FAR is calculated by dividing the number of spoofing attacks by the number of times they were wrongly accepted. Its equation is in Table 2.2 (1). In contrast, FRR represents the ratio of incorrectly rejected real accesses as in Table 2.2 (2).

In face anti-spoofing, a threshold is a value that is used to decide whether a sample is classified as genuine or an attack. By varying the threshold, different operating points can be achieved, and the performance of a face anti-spoofing system can be evaluated based on a range of thresholds. The Receiver Operating Characteristic (ROC) curve is a tool utilised to evaluate the performance of a face anti-spoofing system. The ROC curve plots the FNR versus the TPR for a range of thresholds. The area under the ROC curve (AUC) is a popular performance metric that indicates how well a face anti-spoofing system can differentiate between genuine and attack samples.

Through the development set, predetermining a threshold can be done so that the FRR and the FAR are equal, as in Table 2.2 (3), to produce an Equal Error Rate (EER). The EER point is considered an essential operating point since it indicates the threshold value where the face anti-spoofing system is equally likely to make a false acceptance error as it is to create a false rejection error. Anjoset et al. [48] proposed a Half Total Error Rate (HTER) to compare various FAS approaches in

2011. This is found by calculating the average of FRR and FAR as in Table 2.2 (4).

Other performance metrics are commonly used in face anti-spoofing, which are reported utilising the metrics defined in the standardised ISO/IEC 30107-3 metrics [50], such as the Attack Presentation Classification Error Rate (APCER), Bonafide Presentation Classification Error Rate (BPCER), and Average Classification Error Rate (ACER). APCER is the error rate when the attack presentations are classified as genuine. It is calculated by dividing the number of attack presentations organised as genuine by the total number of attack presentations for a given attack instrument species (PAIS) as in Table 2.2 (5). BPCER is the error rate when genuine presentations are classified as attack presentations. It is calculated by dividing the number of genuine presentations classified as attacks by the total number of genuine presentations, as shown in Table 2.2 (6). ACER is the average of the APCER and the BPCER as in Table 2.2 (7), and it is used at the end to compare the performance of various methods. ACPER and BCPER are threshold-based values, and ACER depends on a threshold τ [50]. These metrics are evaluated at specific operating points corresponding to particular threshold values. For example, APCER and BPCER can be assessed using the Equal Error Rate (EER), the threshold value where the FAR equals the FRR.

In formulas (5, 6, 7), the variables represent the following: N_{PAIS_i} represents the number of attack presentations for a specific Presentation Attack Instrument Subtype (PAIS) denoted as $PAIS_i$. Also, N_{BF} represents the total number of genuine (bona fide) presentations. RES_i takes on either the value 0 or 1, depending on the classification of the ith presentation:

- $RES_i = 0$: Indicates that the ith presentation is classified as genuine (bona fide).
- $RES_i = 1$: Indicates that the ith presentation is classified as an attack presentation.

It's important to note that N_{PAIS_i} specifically relates to the number of attack pre-

sentations associated with a particular PAIS subtype $PAIS_i$. Additionally, RES_i signifies the classification result for each individual presentation, with 0 representing genuine and 1 representing an attack presentation.

Overall, the choice of threshold is important for evaluating the performance of a face anti-spoofing system, and different thresholds can provide additional insights into the strengths and weaknesses of the system.

2.8 Evaluation protocols

We use two protocols to evaluate the discrimination and generalization abilities in the FAS models. We look at the metrics commonly used, those that depend on a specific threshold, such as ACER and HTER, and those that integrate over a range of thresholds, such as AUC, and examine the consistency in their use. Regarding the former, we review the common practices in reporting the evaluation thresholds, focusing on the choice of thresholds, and discuss the consequences of the protocol choice for the reported metric.

2.8.1 Intra-dataset Intra-Type Protocol

The intra-testing protocol is a type of evaluation protocol used in face anti-spoofing, which involves testing the FAS ability to distinguish between genuine faces and spoofed faces that are captured and presented using the same sensor or environment. In intra-testing, the system is trained and tested on the same dataset containing genuine and spoofed face samples. The aim is to evaluate the system's ability to detect different types of spoof attacks, such as printed photos, videos, and 3D masks, within the same testing environment. This protocol is particularly useful for evaluating the robustness of a face recognition system to different types of attacks that can be launched within the same environment. Since the training and testing data come from the same datasets, they have the same domain distribution regarding the recording environment and the subjects' behaviour. The examinations are conducted in a way that does not overlap the subjects' identities. For instance, [89] and [90] achieved satisfactory performance by following this evaluation protocol for

their models.

2.8.2 Cross-Dataset Intra-Type Protocol

The cross-Dataset Cross-Type Protocol is used to test how well the FAS model applies to new domains and attack types that haven't been seen before. This protocol determines how well a model can generalise across different datasets. Most of the time, models are trained on one or more datasets (source domains) and then tested on datasets they haven't seen before. For example, most deep models [91] don't do well when they are trained with Replay-Attack and tested on CASIA-MFSD because the lighting and camera resolution change considerably. This protocol is thought to be the most challenging one in the field of preventing face spoofing.

2.9 Conclusion

In this chapter, we explore biometrics, with a specific focus on facial recognition. We introduce the broader field of biometrics and then delve into facial biometrics, highlighting the vulnerabilities and presentation attack types. We analyze state-of-the-art binary classification methods in Face Anti-Spoofing (FAS) to reveal their limitations. Additionally, we discuss common databases used in conjunction with anomaly detection methods. Finally, we present the evaluation protocols that are integral to our research. In the next chapter, we introduce an alternative approach involving anomaly detection methods, offering numerous benefits over traditional binary methods, marking a significant advancement in the field.

Table 2.2: Evaluation metrics and their equations.

	Metrics	Equation
1	FAR	$\frac{FP}{All\ imposters}$
2	FRR	$\frac{FN}{\text{All clients}}$
3	EER	$(FRR = FAR)$
4	HTER	$\frac{FAR+FRR}{2}$
5	APCER	$1 - (\frac{1}{N_{PAIS}}) \sum_{i=1}^{N_{PAIS}} RES_i$
6	BPCER	$\frac{\sum_{i=1}^{N_{BF}} RES_i}{N_{BF}}$
7	ACER	$\frac{APCER+BPCER}{2}\%$

[1] N_{PAIS_i} is variable represents the number of attack presentations for a specific Presentation Attack Instrument Subtype (PAIS) denoted as $PAIS_i$.

[2] N_{BF} is variable represents the total number of genuine (bona fide) presentations.

[3] RES_i is variable takes on either the value 0 or 1, depending on the classification of the ith presentation:.

CHAPTER 3

Face Anti-spoofing Methods Based on Anomaly Detection

This chapter aims to provide a critical overview of the current research in the field, identify gaps in the literature, and set the stage for our study. We commence this chapter by covering the anomaly detection concept. Then, we comprehensively review the existing literature on face anti-spoofing based on anomaly detection. Then, we compare and contrast the findings of these studies and highlight the best outcomes.

3.1 Introduction

There is a wide range of approaches to the problem of face anti-spoofing. On the one hand, some of the most practical approaches recognise that in several critical applications, the issue of face anti-spoofing is inextricably linked to face recognition. Therefore, a system should only be able to detect malicious attacks on subjects already registered in the database. In such approaches, subject-specific detection thresholds will typically be computed and used in the classification [35, 92]. In contrast, face anti-spoofing can also be approached as a general image classification problem. Given any image or video of a face, can a possible presentation attack be

detected?

In the last few years, there has been a shift toward this generalised approach to the problem. This is evidenced by the fact that cross-database validation of the proposed methods is now becoming the norm. That is, the machine learning classifier is trained on one database and tested on another, meaning that there is no prior knowledge, not only of the subjects but even of the properties of the system that captured the training and test samples, nor is there any knowledge of the environmental conditions under which they were captured. We note that the move towards this broader and considerably more challenging approach to face anti-spoofing is concurrent with the recent great successes of deep learning methods, which regularly report almost perfect results on intra-database testing protocols, even when they are tested on multiple databases. Thus, it seems natural that the researchers are turning their attention to a more challenging problem, where indeed, even state-of-the-art deep learning methods frequently fail to generalise to unseen databases.

Within this broader setting of the problem of face anti-spoofing, with very few or no assumptions on how client and imposter samples are produced, anomaly detection seems to be a natural approach, which is rapidly gaining in popularity, now forming one of the main research directions in the area. Anomaly detection is outlier detection, or novelty detection, i.e., the process of detecting data instances that significantly deviate from the majority of the seen instances.

Moreover, unlike other biometric modalities such as fingerprints [93], or irises [94], the client class, that is, images or videos of human faces, is elementary and cheap to create, and there are abundant samples of it that are freely available. In contrast, in all specialised databases for face anti-spoofing, the imposter class is limited in the number of subjects and the diversity of presentation attacks. Thus, a training method using only the client class is a natural choice. Also, some new research indicates that anti-spoofing challenges can be reframed as anomaly detection problems, potentially improving their generalisation capacity.

Arashloo et al. [27] demonstrated that binary classification-based fPAD algorithms do not generalise well to novel attack types. The study found that training a

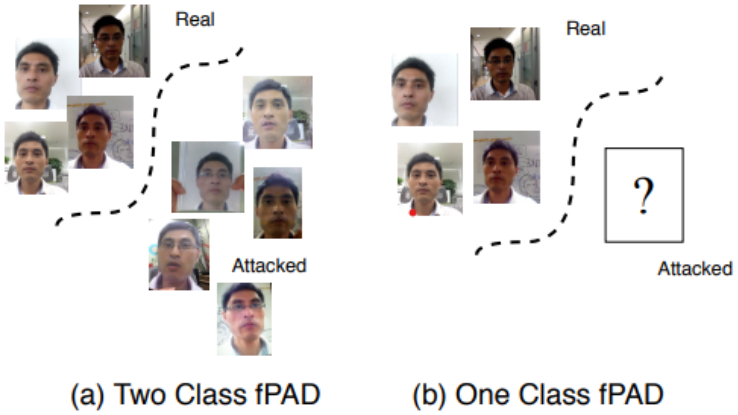

Figure 3.1: (a) (Binary classification) genuine and known attacks are used in training. (b) (One class classification) Only genuine images are available during training [8].

system exclusively on genuine data, as in anomaly detection and only using attack images for model evaluation yields superior generalisation. From [8] Figure 3.1(b) illustrates an anomaly detection-based fPAD, learning spoof detection only from bona fide presentation images of enrolled users.

To review anomaly detection in face-anti-spoofing, we focus on three main aspects: the data and the way they are used, the error metrics and evaluation protocols, and the machine learning algorithms used. We have already mentioned the first two aspects of the data and the evaluation in 2.6, 2.7 and 2.8, respectively. The last aspect is the algorithms; our survey shows that the central theme in the recent and current research is using deep neural networks for feature extraction, followed by a shallow one-class classifier applied to that feature space. We also observe several notable exceptions, the characteristics of which are discussed. We believe that these three aspects are all critical for any further progress in the field.

The rest of the chapter is organised as follows. Section 3.2 presents the existing methods for detecting anomalies (spoofing) in face anti-spoofing. Then, in Section 3.3, we provide the main findings of these anomaly methods, emphasising comparing them to binary methods. In the end, we summarise the main points presented in the chapter in Section 3.4.

3.2 Anomaly detection in Face Anti-spoofing

Anomaly detection identifies unexpected items, events, or observations that deviate significantly from the norm [95]. Anomalies are also called outliers, abnormalities, discordant samples, or deviants [96]. Various research communities have developed diverse approaches for the design of architectures to train anomaly detection systems [97, 98]. Anomaly detection has gained traction in recent years for detecting unseen attacks. When tackling a face anti-spoofing problem, genuine images are considered normal samples, while potential attacks are considered anomalous. According to the most recent studies [99], anomaly-based face spoofing detectors are a promising alternative to their multi-class counterparts.

The genuine class has been observed to cluster closely together and exhibit less variation in feature distribution. Contrarily, attacks can have wide-ranging forms and characteristics. The variations in attacks can lead to anomalies or irregularities in the feature space, which can be detected by using anomaly detection methods.

Anomaly detection methods have improved the generalisation with which genuine faces are classified by taking advantage of the close cluster behaviour of legitimate samples in the feature space. These methods assume that attacks correspond to samples outside the normal distribution's margin. Thus, unseen attacks can be detected because the normal face sample belongs to a well-defined class. The existing face anti-spoofing methods based on anomaly detection are categorized according to two main characteristics: the features, which can either be handcrafted or obtained through deep learning and the type of a one-class classifier.

3.2.1 Handcrafted-Features with One Class Classification

Creating comprehensive face spoofing databases can be difficult because of the wide range of potential attacks and the challenges of gathering sufficient data to tackle all these attack types under varying real-world conditions. Additionally, issues such as imaging sensor interoperability and other environmental factors further complicate the problem. So, the significant challenge in current face anti-spoofing methods is the inability to effectively handle new attacks, resulting in a lack of gen-

eralisation. One of the advantages of employing a one-class method approach is that it requires only relevant information from genuine presentations. In the first work in this field, Arashloo et al. [27] developed a novel strategy for detecting face spoofing on 20 different two-class and one-class configurations. They examined three different dynamic texture descriptors and quality-based features: a local binary pattern operator on three orthogonal planes (LBP-TOP), local phase quantisation on three orthogonal planes (LPQ-TOP); and binarised statistical image features on three orthogonal planes (BSIF-TOP). Then, they used the one class SVM (OC-SVM) [100] and a sparse representation model (OC-SRC) as outlier classifiers. The best result here on Replay-Attack DB was with SRC1+BSIF(w), which was the AUC = 98.98%

Furthermore, [101] categorised the one class classifications (OCCs) as Generative and Non-Generative. An example of generative OCC is the Gaussian Mixture Mode (GMM) with only one client class, while OCSVM is an excellent example of a non-generative OCC. The work by Nikisins et al. in [102] starts with a preprocessor where faces are cropped and normalized, and only clients' images are used in the training set. Technically, the IQM from [103] and [58] was used as a discriminative feature extractor and the one class GMM as a classifier to represent the probability distribution. They aggregated three face spoofing databases, Replay-Attack [6], MSU MFSD [104] and Replay-Mobile [105], to address the generalization problem. Their suggested IQM with one-class GMM generalizes better for the aggregated database's photo-photo-video and video-video-photo protocols, with HTER 14.5% and 29.8%, respectively, which outperforms the OC-SVM performance.

Based on the reviewed papers above, it can be concluded that hand-crafted features alone are inadequate for face anti-spoofing, including anomaly detection. Consequently, there has been a notable emergence of papers exploring deep-learning methods for anomaly detection in face anti-spoofing.

3.2.2 Deep Learning with One Class-classification

When it comes to face anti-spoofing based on anomaly methods that utilise deep learning, they can generally be divided into two categories: those that incorporate transfer learning and those that do not. Transfer learning refers to a technique

where a pre-trained neural network model is utilised as a starting point for a new model, allowing for the transfer of learned features from the pre-existing model to the new one. On the other hand, independent deep learning methods refer to those that do not rely on the pre-existing models and instead train their neural networks from scratch. Thus, face anti-spoofing methods with deep learning can be classified into these two groups based on their utilisation of transfer learning.

Transfer Learning

Due to the fact that deep learning has proven to be effective in solving PA problems, PAD methods based on transfer learning have shown promising results in particular. Overfitting can be avoided with transfer learning when there is limited training data. Transfer learning also saves computational resources since it does not start from scratch. Accordingly, many works emerged that used transfer learning based on anomaly detection [8, 35, 92, 106–108]. All of these works utilized these pre-trained models as feature extractors. From a practical viewpoint, they improved the performance compared to the methods with hand-crafted features.

Baweja et al. [8] work was inspired by Oza and Patel in [109]. The later network uses a one-class CNN (OC-CNN) technique to get over the older OC-SVM's restrictions, comprising a feature extractor network and a classifier network. Baweja et al. [8] used the end-to-end fPAD model, where the classifier and learning features are learned together. The authors used the pre-train VGG16 network, omitting the softmax regression layer, and compared their OC-CNN work to four baseline methods: OCSVM, SVDD, MD, and OCGMM. They used a Gaussian Distribution to model a "pseudo-negative" class, where a weighted running mean found the mean. They used the APCER, BPCER and ACER as evaluation metrics, and their proposed method resulted in values of 25.047, 16.539, and 20.739, respectively.

Client-Specific Approaches

Client-specific information is utilised to determine a unique threshold for each client, subsequently used for decision-making during testing. This threshold is based on subject-specific score distributions. The first work recognising the advantage of

using client-specific information was [110]. Based on the client-specific thresholds, the works of [35, 92, 106, 107] all depend on client-specific thresholds to determine the classification outcome.

Throughout the work of Fatemifar et al., [106], each frame was photometrically normalized using the retina method [111] to minimize the variation in lighting conditions. They proposed FAS models and adopted client-specific thresholds for each one. Four of the common pre-trained deep neural networks were used; GoogleNet [112], ResNet50 [113], VGG-verydeep-16 [114], and VGG-Face [115]. The output feature vectors were passed to one of the four common anomaly detectors; One-Class SVM, One-Class SRC, One-Class MD, and One-Class GMM. Their work concluded that an enhancement for detector performance is possible using client-specific detection thresholds instead of a global threshold. The HTER result for the Replay-Attack DB is in Table 3.1 and is 2.82. the Replay-Mobile DB achieved the best result with GoogleNet + MD, which is 13.70, and the Rose-Youtu received the best result with ResNet50 + GMM, namely 13.6. All previous results used client-specific thresholds; the performance degrades when using global thresholds. Overall, [106] showed that the results from the suggested model outperform the class-independent formulation and traditional binary classification models in a promising manner.

Fatemifar et al. [92] proposed a novel ensemble method for fusing one-class classifiers, which are One-class SVM, One-class MD, and One-class GMM classifiers, as well as using common CNNs (GoogleNet, ResNet50, and VGG-verydeep-16 (VD16)) for deep feature extraction.

To mitigate the influence of environmental lighting variations, they employed a strategy of dividing the original face image into seven regions and taking into account weighted averages. These seven face regions, along with three one-class classifiers and three convolutional neural networks (CNNs), collectively constitute a set of 63 distinct spoofing detectors. Utilising Genetic Algorithms (GA), they create a fusion approach based on a weighted average of component classifiers. Their method was validated by extensive trials on three spoofing datasets: Replay-Mobile, Replay-Attack, and Rose-Youtu protocols. Also, their model's results outperformed the

class-independent formulations and multiclass classification techniques.

Following [92] and [106], which showed the One-class GMM as the most competent classifier among all others, the HTER result in [92] with the GA in the Replay-Attack DB is 1.43, which is better than [106]. Also, the HTER results for Replay-Mobile and Rose-Youtu DBs are 9.95 and 9.30, which are again better than [106].

The authors in [107] proposed a stacking ensemble approach [116] based on anomaly detection and client-specific. Also, they offered a novel GA optimisation with two steps. In the first step, a novel anomaly-based fitness function called Proportion Above Unity (PAU), which does not need a-prior information about spoofing, was used to measure the competency of the POC. Then, they pruned the stacking ensemble using the GA and OCCs. In the second step, they retrained the pruned OCC using GA to build a final stacking ensemble. Again, the databases used in the previous works [92] and [106] are used here. Even though the study's main goal was to see how well anomaly-based methods could work, not how well they could beat two-class methods, the results of the experiments showed that the suggested client-specific anomaly-based Stacking method perform very well when compared to two-class methods.

Fatemifar et al.'s [35] proposed approach combines both generative and discriminative models to enhance the accuracy of the PAD system. Generative models capture the statistical characteristics of genuine samples, while discriminative models aim to differentiate between genuine and spoofed faces. Instead of using a generic classifier that applies to all subjects, they develop client-specific one-class classifiers. This work involves training individual classifiers for each person in the dataset. Capitalizing on the effectiveness of deep neural networks, notably Convolutional Neural Networks (CNNs) such as GoogleNet, ResNet, VGG16, and VGGFac, they employed these networks primarily as feature extractors.

Then, they used the score distributions of each client-specific threshold to differentiate between genuine and spoofed faces. The experimental results and evaluations demonstrate the effectiveness of the proposed method in accurately detecting presentation attacks in face recognition systems.

The performance of one-class spoofing detection models can be enhanced by using client-specific information during model construction and decision threshold setting. [117] formulated the FAS problem as a one-class kernel Fisher null-space regression anomaly detection problem. Their work showed that even under the challenging conditions of "unseen", state-of-the-art performance can be achieved with generic pre-trained deep CNN models. For increased detection performance, they presented a multiple kernel fusion anomaly detection technique that integrates complementary information from different perspectives of the problem.

To improve the performance of the anomaly detection, [90] fused OCCs using Weighted Averaging (WA) and client-specific information in two phases; Particle Swarm Optimisation (PSO) [118] and the Pattern Search (PS) [119] algorithms were used to avert the local minimum problem and to enhance the generalisation capability of the WA fusion. Also, they presented an innovative scoring normalisation method to normalise the heavy-tailed distributions. Within their work, three CNNs feature extractors and two anomaly detectors (GMM and MD) were employed. They experimentally demonstrated that the proposed client-specific anomaly-based fusion obtains outstanding results in the unseen domain. Fatemifar *et al.* in [120] aimed to evaluate the effectiveness of using WA fusion to combine multiple anomaly classifiers based on genuine-access data. A novel three-stage optimization approach was proposed to optimize the parameters of WA, including a hybrid optimization method using both (GA) and Pattern Search (PS) to improve the exploration of the weight space. Second, a novel two-sided score normalization method to enhance anomaly detection performance, And third, an ensemble pruning method to improve generalization performance. Furthermore, the proposed model was trained using client-specific information to improve the anomaly detection ensemble further. The experimental results revealed that the introduced WA fusion approach outperformed the state-of-the-art anomaly-based and multi-class methods.

Table 3.1 provides a summary of how various methods perform in distinguishing real faces from fake ones across different face databases, including Replay-Attack, Replay-Mobile, and Rose-Youtu. The table lists different methods used to spot fake faces, such as ResNet50 + MD, WA, GA-Stacking, and others. These methods

were tested on different face databases, each containing real and fake faces. The table shows scores (HTER values) for how well each method performed. Lower scores mean better performance. GA-Stacking stood out as the top performer, particularly on the Replay-Mobile and Rose-Youtu databases. On the Replay-Attack database, some methods achieved perfect scores, like kernel-regression, indicating excellent performance. However, most methods faced greater challenges and had higher scores on the Replay-Mobile and Rose-Youtu databases, suggesting a tougher time distinguishing real and fake faces in those scenarios.

Table 3.1 displays the HTER results for the methods mentioned: [35, 90, 92, 106, 107, 117, 120]. Among all the methods mentioned, the GA-Stacking method [107] achieved the best results for both the Replay-Mobile DB and Rose-Youtu regarding HTER. The HTER values on the Replay-Attack DB were generally 0 with [90, 107, 117, 120], indicating good performance on this dataset. On the other hand, the HTER values on the Replay-Mobile DB and Rose-Youtu were worse, suggesting that the methods may have faced more challenges or exhibited higher error rates on these databases.

Table 3.1: HTER (↓) results of client-specific methods for Replay-Mobile and Rose-Youtu DB

Method	**Replay-Attack**	**Replay-Mobile**	**Rose-Youtu**
ResNet50 + MD [106]	2.82	-	-
WA [92]	1.43	9.95	9.30
GA-Stacking [107]	**0**	**3.75**	**3.61**
kernel-regression [117]	**0**	13.6	-
GoogLeNet + MD [35]	7.84	**-**	**-**
WA(PSO+PS) [90]	**0**	5.85	5.65
WA(GA+MMS+PS) [120]	**0**	5.35	5.12

Deep Learning Methods

Recent studies have highlighted that the current approach, which involves using CNNs to extract features, needs to generalise more effectively. The problem of generalised presentation attack detection is one of the most significant obstacles that must be overcome to robust the anti-spoofing technologies in actual contexts.

In [121], an anomaly detection system based on CNNs was used to detect face presentation attacks (PAD) using multi-channel images. Multi-channel images provide valuable information to differentiate between various types of attacks, and the anomaly detection method allows for effective generalisation. Two methods were used in this study; the first method involved a deep autoencoder, which is a multi-channel convolutional autoencoder (MCCAE) network that extracts the common factors from bonafide samples and is trained to minimise the mean-square error (MSE). When tested on the WMCA DB, the result was a 0.9961 AUC value. The second method used deep support vector data description (SVDD), with the encoder part of the MCCAE network being used to extract features for the SVDD, resulting in an AUC value of 0.9393. Although the deep SVDD had a lower computational complexity, it performed somewhat worse than MCCAE. The test results indicated that an intermediate feature representation could express the differences between the distributions of genuine photos and most attacks in feature space. Therefore, it supports the hypothesis that all genuine data have the same inherent characteristics.

Arashloo [122] proposed a multi-kernel learning (MKL) algorithm based on the Fisher one-class null method and expressed the problem as an optimisation challenge based on the saddle point. They developed an effective approach to address the saddle point optimisation problem linked to the proposed one-class MKL algorithm. This method involved utilising different deep Convolutional Neural Network (CNN) representations as kernels and introducing MKL algorithms incorporating an (r, p)-norm matrix regularisation constraint. They evaluated the algorithm on general object images and FAS DBs in an unseen attack scenario. They compared its performance against the baseline and other methods from the literature, with multiple kernels and deep one-class end-to-end learning approaches. The performance of the suggested method was better than the current methods.

Table 3.2: AUC (%) and HTER results of intra-database test on the Replay-Attack DB using Deep learning methods.

Method	**AUC**(↑)	**HTER**(↓)
Live Correlation Loss [89]	90.52	-
Localised MKL [99]	100	0
Matrix [122]	100	0
Smi-supervised [9]	**-**	3.5

To increase the generalization capability of the attention layers, [123] proposed an attention auto-Encoder (AAE) based one-class model with only bonafide images in training. The attention-based model can exclude irrelevant information and concentrate on identifying distinctive characteristics of the real face. They used the reconstruction error and the centre loss from the latent layer of the AAE network to calculate the spoofed score. The findings demonstrated that their approach is better at distinguishing between genuine and spoof faces when presented with unique attack patterns that were not part of the training set. Chen et al. [9] proposed training the model solely with genuine face data, using a convolutional Encoder-Decoder network as a generator and another convolutional network as a discriminator. The generator and the discriminator are trained by working together and competing against each other to understand genuine faces better. They took the optical flow maps derived from video frames as their entry point.

The results for [9] and [123] on Replay-Attack and CASIA-MFSD with the cross-database protocol are shown in Table 3.3.

A novel approach to face anti-spoofing based on one class using bonafide images alone learning with live correlation loss was proposed in [89]. Specifically, encoder-decoder networks were initially trained with live faces only to extract latent features. These features could compactly represent various live facial properties in the embedding space and produce spoofing cues, which are easily obtained by subtracting the original RGB image from the generated one.

The work in [99] investigated a particular aspect of the face presentation attack detection problem, focusing on how MKL in a one-class setting can benefit from

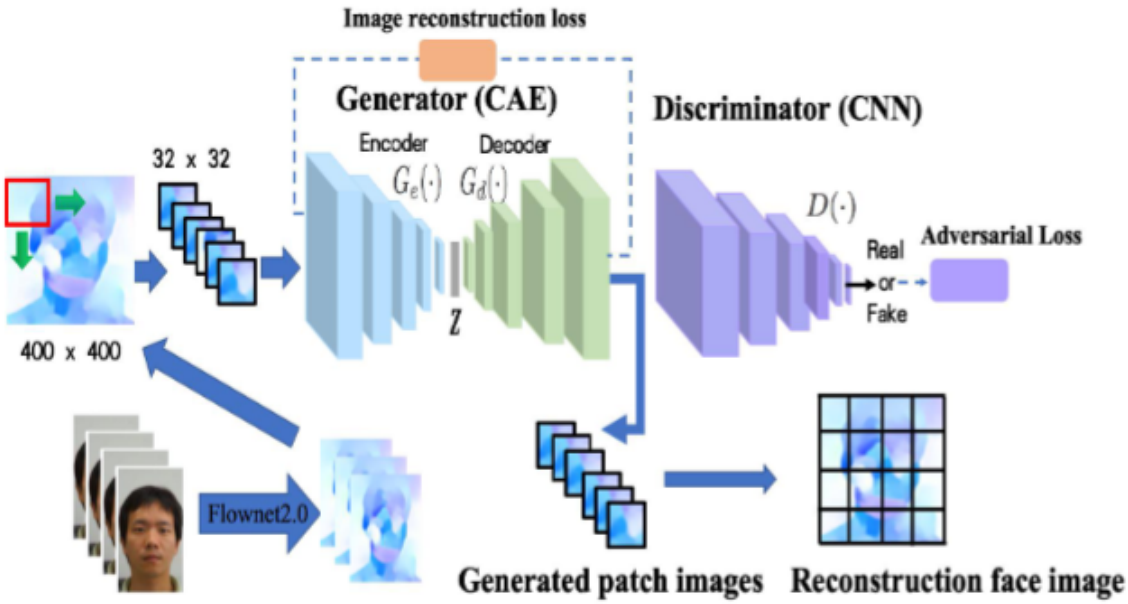

Figure 3.2: The proposed method in [9] consists of a generator and a discriminator trained to understand the underlying structure in the live faces data.

the intrinsic local structure present in genuine face samples. The authors proposed a localised multiple kernel learning approach as a convex optimization problem for the pure one-class classification of attacks involving unknown/unseen face presentation. Furthermore, they conducted a theoretical analysis to assess the generalization performance of the approach.

Table 3.2 shows the results for all these works. We can note that the best results are the methods with MKL [99,122], which are equal to HTER 0. This approach has limitations, such as a high computational cost due to using multiple kernels, which may limit its applicability in real-world scenarios. Furthermore, it is evaluated on general object image datasets, which may not fully represent the complexity and diversity of face presentation attacks.

3.3 Summary of The Main Findings

The binary classification methods in FAS aim to distinguish between clients and imposters using a pre-defined set of features. In contrast, anomaly detection methods aim to identify imposters by detecting patterns that deviate from the normal distribution of the clients. Binary classification methods may achieve higher accuracy in

identifying spoofed faces when compared to anomaly detection methods, especially when the spoofing attacks are well-defined and the features used for classification are effective. However, anomaly detection methods are better at detecting new types of spoofing attacks not included in the training dataset. In addition, binary classification methods are vulnerable to adversarial attacks, where an attacker can manipulate the input image to bypass the classification model. Anomaly detection methods are more robust to such attacks, as they rely on detecting patterns that deviate from the normal distribution, which may be harder to manipulate.

Table 3.3: Comparison of face anti-spoofing methods on cross-database testing CASIA-MFSD (C)and Replay-Attack datasets (RA). The top four rows of HTER results correspond to binary methods, and the bottom two rows are the one-class methods.

Method	**C/RA**	**RA/C**
Auxiliary 2018 [124]	27.6	28.4
STASN 2019 [60]	31.5	30.9
CDCN 2020 [91]	15.5	32.6
CDCN++ 2020 [91]	6.5	29.8
One-class-attention 2021 [123]	20.0	26.9
Smi-supervised 2022 [9]	15.6	44.1

Table 3.3 provides a summary of the performance of various face anti-spoofing methods when tested on different databases, namely CASIA-MFSD(C) and Replay-Attack (RA). The evaluation metric used is the Half Total Error Rate (HTER). The table includes results for both binary methods (top four rows) and one-class methods (bottom two rows). Interestingly, the results indicate that the one-class classification methods perform well under the cross-database testing protocol. These findings highlight that we can see that anomaly detection methods are the promising direction to work within the face anti-spoofing domain.

3.4 Conclusion

Due to the diversity of spoofing attack data, the conventional two-class technique presents several challenges, including the difficulty of establishing an efficient decision boundary for classification. Furthermore, acquiring attack data is complex and time-consuming and cannot cover all possible unforeseen attacks. While increasing the size of genuine access data is possible, doing so will result in an imbalanced training set.

Therefore, two-class systems might be unable to generalise or adapt to new attacks. One-class methods provide significant advantages. By utilising only genuine data for training, it exhibits robustness against the negative effects of spoofing data diversity. Furthermore, extending the training set becomes easier in one-class methods. As a result, anomaly-based approaches are better equipped to detect unseen and entirely novel attack samples.

CHAPTER 4

Colour Processing in Adversarial Attacks on Face Anti-spoofing

In many typical daily applications, face recognition systems are frequently utilised for user authentication. Nevertheless, their susceptibility to basic presentation photo attacks, like when an imposter gains entry to a system by presenting a photo of the legitimate user in front of the camera, imposes limitations on their deployment to contexts where security is not deemed paramount. Face anti-spoofing methods are countermeasures against such attacks, aiming to confirm the presence of a genuine client's live face, rather than a mere photo, in front of the camera of the face recognition system.

In this Chapter, we examine the effectiveness of a face anti-spoofing method against imposter photo attacks, with the added assumption that the images used in the attack may have been manipulated, and in particular, we introduce a novel method for generating adversarial images that can successfully fool CNN-based anti-spoofing. We employ one of the most well-known CNNs, the ResNet50 [125], which is evaluated against imposter attacks with unprocessed photos and attacks with processed photos. The latter had their saturation increased, making them look more lively. The assumption is that the class of processed photos will be closer to the class of client images, and thus, it will be more challenging for the algorithm to

differentiate between these two classes.

4.1 Introduction

Face recognition [17] has long been established as the biometric method of choice for everyday applications, such as mobile phone or PC login. However, its use in security-critical applications is currently restricted to controlled environments, such as airport passport control, but not, for example, money withdrawal from street ATM machines.

The main reason behind this limitation is that face recognition is considered particularly vulnerable to *presentation attacks*, where one may gain access by presenting in front of the system's camera a photo or a video of the user they impersonate [126]. Developed as countermeasures to such attacks, *liveness tests* are binary classifiers aiming at distinguishing between the genuine *client* and illegitimate *imposter* images or videos. In this Chapter, we aim to answer our first question which is **RQ1**: By what methods can we enhance the effectiveness of a certain type of presentation attack, and by how much does the accuracy of the face anti-spoofing test drop?

As a result, we study *adversarial attacks* on face anti-spoofing based on deep neural networks. In particular, we study the extent to which increasing the saturation of an imposter face image degrades the ability of the neural network to classify it correctly. Our approach was motivated by the observation that, generally, the client images have more vivid colours than the imposter ones. This divergence in color vibrancy results from the favorable conditions under which client images are taken, which preserve the richness of their colors. In contrast, imposter images often lose color intensity during the process of recapturing, leading to a reduction in their overall vibrancy [127].

We note that the study of adversarial attacks on machine learning classifiers is the focus of a large body of recent research and is considered one of the main methodologies for understanding and improving neural network performance. However, to the best of our knowledge adversarial attacks on face anti-spoofing have not been studied before our work, with the exception of [32, 128], where however

adversarial attacks on traditional only machine learning methods are studied, which is not satisfactory since deep neural networks have established themselves as the state-of-the-art in almost every machine learning task.

In the context of classifiers for the face anti-spoofing tests, beyond the issue of understanding and improving the classifier, another question we want to address in this Chapter, **RQ2**: is whether the proposed adversarial attack can be converted into a direct presentation attack? That is, we want to verify that the same performance degradation will be observed if instead of just manipulating the imposter images of the database, we create new imposter images by increasing the saturation of client images, printing them or displaying them on an electronic device and capture an image of them. In other words, we would like to verify that the all-digital adversarial attack on the test database of the classifier can be converted into a physical attack on a real-face anti-spoofing test system.

As the execution of that physical attack is a labour-intensive process, we run a limited experiment, which, however, gives a clear indication that the corresponding presentation attack is enhanced by the manipulation of the client images before presenting them to the system's camera. This was not an unexpected result, since we had already established that saturation increases lead to classifier performance degradation, and we naturally expect that presenting to the camera a higher saturation image will also result in a higher saturation image as the camera's output.

Contributions: We propose a colour manipulation adversarial attack to a face anti-spoofing based on a deep neural network. To the best of our knowledge, it is the first study of adversarial attacks on deep neural networks in the context of face anti-spoofing detection. In a second contribution, we conducted an experiment the result of which indicates that the proposed adversarial attack can be converted into a direct presentation attack.

The rest of the Chapter is organised as follows. In Section 4.2, we review the related work. In Section 4.3, we explain the proposed methods in this Chapter to generate novel attacks. In Section 4.4, we present the proposed face anti-spoofing method to evaluate it. In Section 4.5, we present the results of this chapter. Then, we briefly conclude in Section 4.6.

4.2 Related work

In this section, our focus is on discussing the related work that specifically pertains to the topic of adversarial attacks in face anti-spoofing. While Chapter 2 provided a general overview of the broader field of face anti-spoofing, we now narrow our focus to studies and research specifically related to adversarial attacks in this domain.

4.2.1 Adversarial Attacks

Adversarial attacks can be categorized into two main types: white-box attacks and black-box attacks. White-box attacks [129] involve generating adversarial perturbations by utilizing the gradient information of the targeted model. In other words, the attacker has complete knowledge of the model's architecture, parameters, and training data. This allows them to calculate the gradients of their responses and optimize the perturbations of the instrument of attack to maximize the effect on the model's response. However, in our context, the most relevant adversarial attacks are the black-box [130] ones, where the attacker does not have access to the hidden layers of the network or, more generally, any information about the type and the parameters of the classification algorithm. Black-box attacks rely on crafting adversarial examples using a substitute model, and this type of attack is more challenging as they require finding vulnerabilities in the model without direct knowledge of its internals.

Adversarial attacks have been extensively studied in the related field of face recognition, and they can often be very simple in nature. In [131], the authors generated attacks by adding a small perturbation to a single pixel, or a small set of pixels. We also note that there are several open-source software tools for generating adversarial images, e.g. DeepFool [132]. There is a growing concern regarding the proliferation of malicious digital manipulation attacks targeting face videos. Such attacks include techniques like morphing attacks using generative models like StyleGAN. Morphing attacks [133] involved blending or morphing multiple faces together to create a single composite face. Generative models like StyleGAN have been used to generate high-quality and visually appealing morphed faces. These

attacks aim to create a face that combines the features of two or more individuals, leading to identity ambiguity and potential misuse.

Only a limited number of studies have studied attacks in the physical domain, commonly called physical domain attacks. One notable example is the work by Sharif et al. [134], where they attacked a face recognition system by printing malicious examples on eye glasses frames. By wearing such glasses, an attacker could successfully impersonate another person. The objective of their research was to find an adversarial perturbation that could deceive the face recognition system for a large class of images. To achieve this, they employed geometric transformations, commonly encountered in the recapture process, on a collection of portrait photos gathered by the attacker.

Regarding physical adversarial attacks in face anti-spoofing, there is a limited number of works on that, especially in conjunction with deep learning. In [32, 128], they evaluated the face anti-spoofing method by Tan et al [5] which used Logistic Regression as a binary classifier under different attackers' assumptions (different amounts of sharpening of the imposter images). They found that sharpening the imposter images decreases the accuracy of the Tan et al [5] method while smoothing, or sharpening followed by smoothing, improved the accuracy. In [135], they proposed a method for crafting adversarial images that exploit the weaknesses of CNN-based anti-spoofing techniques. By carefully manipulating the input images, they created adversarial examples that are capable of bypassing the anti-spoofing measures and successfully tricking the face recognition system. In [136], they proposed a framework that exposes the fine-grained adversarial vulnerability of face anti-spoofing models. They introduced a model comprised of a multitask module and the semantic feature augmentation (SFA) module, which incorporates a data distribution prior to generating adversarial examples that exploit more discrimination-related gradient directions. The effectiveness of the SFA module was quantitatively measured by the increase in attack success rate.

4.3 Methods

In this section, we will provide a concise overview of the background of the methods used in this chapter. The aim is to offer a brief explanation of the underlying concepts and principles behind these methods, providing the necessary context for better comprehension.

4.3.1 ResNet

Currently, Deep Learning techniques are being used in almost every academic field. CNNs are the primary technique for deep learning. Transfer learning is a powerful tool in deep learning that allows us to leverage the knowledge learned by pre-trained models and improve performance on new tasks with limited data and the pre-trained model has already learned many useful features from the large dataset it was trained on. In particular, the acquisition of discriminative and intrinsic feature representations in pre-training can be crucial in developing a reliable FAS algorithm. There have been several studies that have demonstrated the usefulness of transfer learning for FAS. [137].

Residual Network, also known as ResNet, is a type of deep neural network architec-

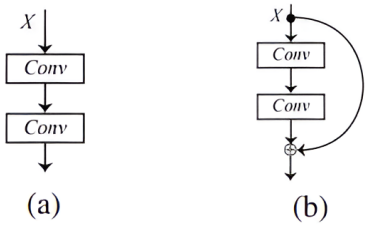

Figure 4.1: (a) a standard CNN (b) a Residual Block in ResNet.

ture introduced by Kaiming He, et al. [125], and it is widely used in computer vision tasks, particularly image classification. ResNet has been the winner of the classification task in the 2015 ImageNet Large-Scale Visual Recognition Challenge [138], reaching a 3.57 % error. ResNet is available at different depths, such as ResNet-18 and ResNet-50. The main idea behind ResNet is to allow for the training of very deep neural networks by addressing the problem of vanishing gradients [139], which

occurs in deep networks when the gradients become very small and cannot effectively propagate through the network during backpropagation. As shown in Figure 4.1, the residual connections in ResNet allow for the creation of very deep networks by bypassing one or more layers and passing information directly to later layers. This helps to prevent the vanishing gradients problem and makes it possible to train much deeper networks than was previously possible.

4.3.2 Color Manipulation Technique

The color manipulation technique used here modifies the saturation component in the HSV colour space by different amounts to create new attacks. The HSV color space [140], is a colour model that represents colours based on their perceived attributes of hue, saturation, and value. The hue and saturation dimensions of an image represented in the HSV colour space are responsible for determining its chrominance, whereas the value dimension is responsible for determining its brightness [3]. Saturation is a measure of the purity and intensity of the colors in an image. When the saturation of an image is increased, the colors become more vivid. Modifying the saturation component in an HSV image can produce different effects, such as making the image appear more vibrant by increasing the saturation, or more muted by decreasing it.

4.4 Experimental setup

As we mention in Section4.3 the FAS classifier that we use is based on ResNet50 which has been evaluated against common face anti-spoofing attacks in [59]. For training and testing, we used the Replay-Attack database [141].

4.4.1 Implementation and training

All code was written in Python, on the Pytorch deep learning platform, and the experiments ran on an Intel Core i7 CPU 64GB RAM PC. We used the pre-trained convolutional part of ResNet50 and trained with our images for 24 epochs, using the Adam optimizer with a learning rate of 0.0001, while the batch size was set to

20. The custom classifier as in the Figure 4.2 that contains a fully connected layer with ReLU activation followed by a Dropout with a 20% chance of dropping and a fully connected layer with log softmax output.

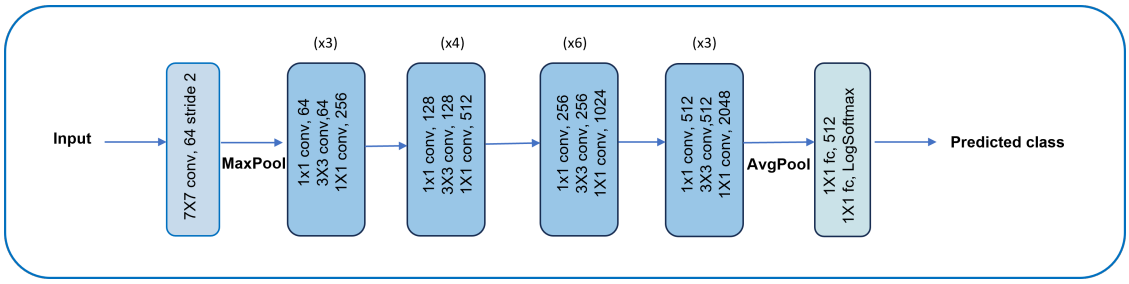

Figure 4.2: ResNet-50 neural network architecture with adding the custom classifier layers.

4.4.2 Validation

Using a within-subject validation protocol, we trained ResNet50 with 1279 images from 14 subjects. The test set consisted of 290 clients and 310 imposter images, from all 14 subjects. On the clients, we obtained a True Negative Rate (TNR) of 99%, while on imposters, a True Positive Rate (TPR) of 98% for a total accuracy rate of 98%.

The impressive performance of ResNet50 under a within-subject protocol masks the difficulty of the liveness classification task on images of previously unseen faces. In the next step, we validated the network under a cross-subject protocol, training ResNet50 on 1082 images from 12 subjects and testing it on 240 images (120 clients and 120 imposters) from 2 different subjects. This time we obtained an 88 % accuracy rate, which indicates the nature of the challenges in cross-subject spoofing detection.

4.5 Results and Discussion

4.5.1 Adversarial Attack

The adversarial attack was validated with the cross-subject validation protocol described in Section 4.3. Modifying the saturation component in an HSV (Hue, Satura-

tion, Value) image can produce various effects depending on the specific adjustments made. Here's a general methodology for modifying the saturation component in an HSV image. Firstly, convert the original RGB (Red, Green, Blue) image into an HSV color space. Secondly, the saturation component in an HSV image represents the intensity or purity of colors. Modifying the saturation can be done by scaling or shifting the values. Increasing saturation will make the colors more vivid, while decreasing it will desaturate the image. To increase saturation, we multiply the saturation values s by a factor α greater than 1. To decrease saturation, we multiply the saturation values s by a factor α less than 1. After adjusting the saturation, ensure that the values remain within the valid range (0 to 255 for the saturation component). we may need to clip values that go beyond this range. Once we made the desired adjustments to the saturation component, convert the modified HSV image back to RGB color space.

$$s \rightarrow \min\{\alpha \cdot s \ , \ 255\}.$$

Figure 4.3 shows how the proposed method was conducted to generate the adversarial attack.

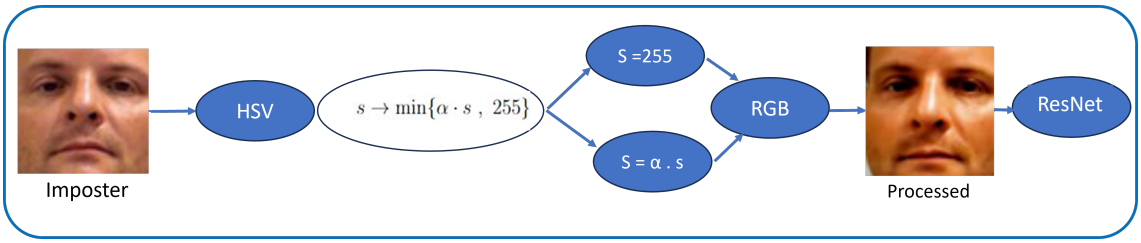

Figure 4.3: Our proposed method to generate the Adversarial Attack.

Also, Figure 4.4 shows one imposter image from each subject undergoing that type of colour manipulation. Table 4.1 provides the achieved True Positive Rates (TPRs), indicating the ratio of correctly identified imposters by the model, alongside the display of the model's average loss. In the experiments we cover a range of values of alpha, from $\alpha < 1$, which decreases saturation, to $\alpha > 1$, which increases it. The main observation is that when $\alpha > 1$ the TPR is lower than that on the original images, which corresponds to $\alpha = 1$, indicating a successfully adversarial attack. In contrast, when $\alpha < 1$, the TPR is higher, providing further evidence

for the effectiveness of the attack. As expected, the average loss values exhibit the opposite pattern. So, we note the significant decrease in the TPR, which, for example, translates into 12% more imposter attacks being successful when the saturation value is multiplied by 1.25. Note that as we do not manipulate the client images, the TNR is the same as in Section 4.3.

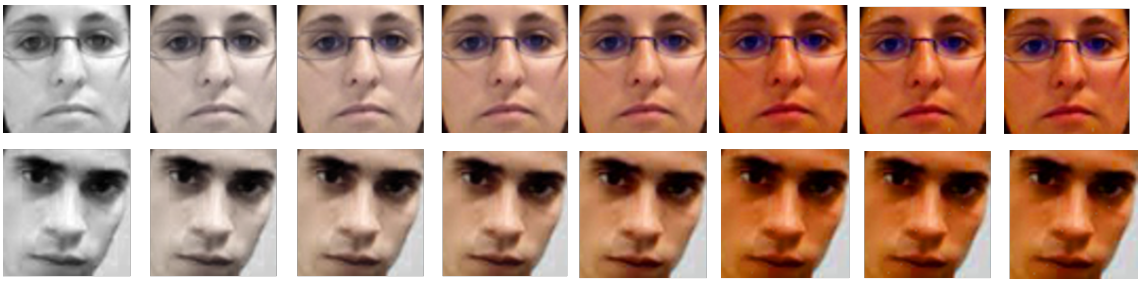

Figure 4.4: Saturation linearly scaled by a constant α and capped to 255. From left to right: $\alpha = 0, 0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75$.

Table 4.1: TPR T_p and average loss L for various values of α. The bold numbers corresponds to the original images.

α	0	0.25	0.5	0.75	**1**	1.25	1.5	1.75
T_p	95	89	99	89	**88**	75	75	77
L	.21	.33	.06	.27	**.38**	.57	.56	.53

In a second experiment, we put the saturation of all pixels of all images to a fixed value s. The corresponding example images are shown in Figure 4.5 and the corresponding TPR and average loss values are shown in Table 4.2. Additionally, it's noticeable that as the saturation increased, the imposter images moved towards a closer resemblance to the client images. In this context, within the saturation range of 63-233, the True Positive Rate (TPR) exhibits a consistent monotonous decline. It commences with a relatively high value of 94% for $s = 63$ and gradually decreases to a minimum of 55% for $s = 223$ The extreme values of $s = 31$ at the left end of the table and $s = 255$ at the right end of the table exhibit different behaviour, in an interesting phenomenon that in the future we would like to study further.

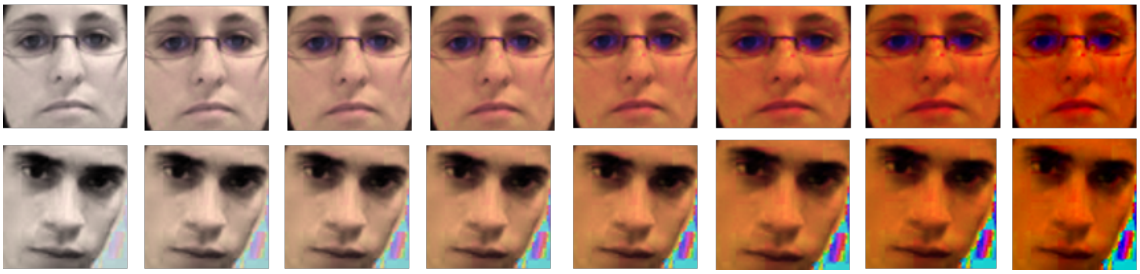

Figure 4.5: Fixed saturation values s. From left to right: $s = 31, 63, 95, 127, 159, 191, 223, 255$.

Table 4.2: TPR Tp and loss L for various fixed values of s

s	31	63	95	127	159	191	223	255
L	.20	.17	.15	.48	.77	.82	.98	.68
T_p	92	94	91	80	70	58	55	73

4.5.2 Presentation Attack

The presentation attack was validated on imposter images created from the client images of the Replay-Attack. Three client images from each of the two subjects were:

i. printed on A4 paper using commodity printer.

ii. displayed on a commodity Lenovo LCD screen with a resolution of 1920 x 1080 pixels.

iii. displayed on an iPhone 10 XS MAX screen with a resolution of 2688 x 1242 pixels.

and then captured with an iPhone camera. The acquired images were manually cropped and resized to 60×60. The whole process was repeated with the saturation of the client images put at a fixed value of 180. Figures 4.7-4.9 show one face image for each subject and each condition.

The corresponding TPR and loss are reported in Table 4.3. We note that, as expected, in all three forms of physical attacks, a high saturation value of $s = 180$ leads to a lower or equal TPR when compared to the corresponding imposter images that were produced by the same physical method from unprocessed client

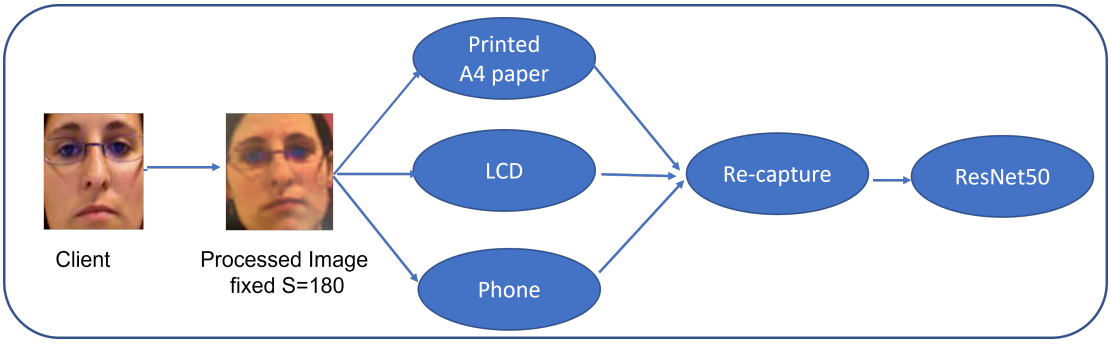

Figure 4.6: Presentation Attack.

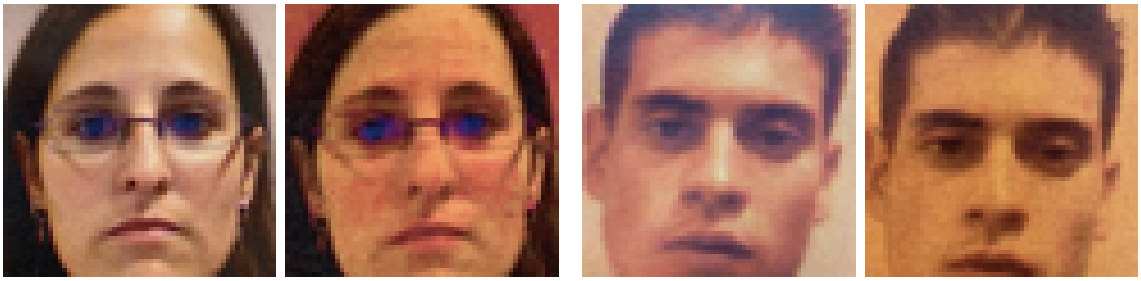

Figure 4.7: Paper print: for each pair of images, the left is a photo of the original client and the right is a photo of the processed client.

images. We note that while the test set is very small for the reported TPR to have significance, the reported loss values provide further evidence for the validity of the conclusions.

Table 4.3 reveals that the Liquid Crystal Display (LCD) attack maintains a TPR of 100 for both unprocessed and processed images, indicating that this type of attack remains unaffected by the color processing. In contrast, the Printed and Phone image attacks exhibit differences in TPR and loss values.

Table 4.3: TPR and (loss) for all 6 types of presentation attacks.

	Original	**Processed**
Paper	100(0.05)	83(0.46)
LCD	100(0.03)	100(0.07)
Mobile	100(0.18)	66(0.42)

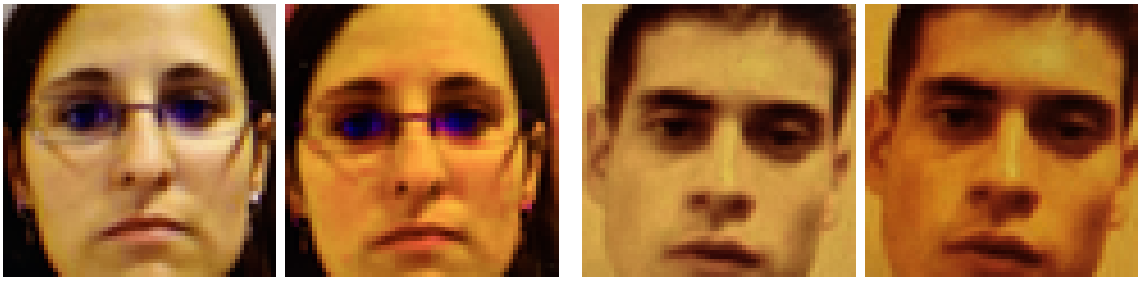

Figure 4.8: LCD: for each pair of images, the left is a photo of the original client and the right a photo of the processed client.

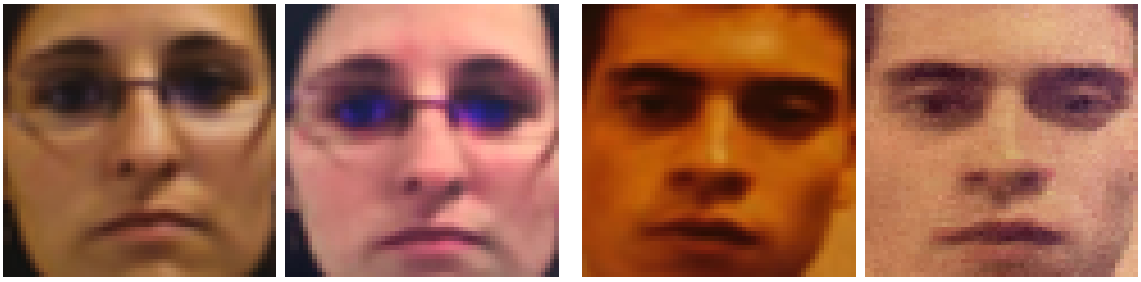

Figure 4.9: iPhone: for each pair of images, the left is a photo of the original client, and the right is a photo of the processed client.

4.6 Conclusions

In real-life situations, we should expect that an attacker will process a client's images before using them. So, we evaluated the deep learning network ResNet50 against attacks in conjunction with certain image processing operations. Our hypothesis was that the model was trained to learn and extract meaningful features from input images, such as color information and texture patterns, with a typical reliance on detecting subtle color differences between genuine and fake face images. Then, when the saturation of an imposter face image is increased, these subtle differences become less prominent, posing a greater challenge for the face anti-spoofing system to accurately identify the image as fake. In particular, as expected, we found when the saturation of an image is increased, the colors become more vivid and that decreases the accuracy of the face anti-spoofing algorithm. Moreover, initial results indicate that this adversarial attack can become the basis for an effective physical presentation attack, in which the imposter increases the saturation of a face image before printing it on paper or displaying it on the screen of an electronic device.

CHAPTER 5

Training Dataset Construction for Anomaly Detection in Face Anti-spoofing

In Chapter 4, we discovered that it is difficult to anticipate all the attackers' usages of instruments of presentation attacks in a real-world scenario. Thus, we realised that it is preferable to employ one-class classification methods that restrict the training sets to genuine images only. In this chapter, we are proposing the enhancement of the performance of face anti-spoofing by augmenting the training dataset with clients' samples harvested from different resources, such as wild images. The process of augmenting the training dataset involves carefully selecting and integrating samples that accurately represent the diverse challenges encountered in practical scenarios. Wild images, for example, capture faces under different lighting conditions, poses, backgrounds, and camera qualities. By including such samples, the model achieves a significant performance improvement under cross-database testing protocols.

5.1 Introduction

Face anti-spoofing is a technique used to detect attempts to fool a face recognition system by presenting a fake face image or video. The main challenge for developing a

robust face anti-spoofing system is a large number of different types of *presentation attacks* the system must learn to recognize. For example, an imposter could be presented to the face recognition system with a printed photo, a screen displaying a still image, or a screen replaying a video. A multitude of other factors, such as the quality of the printed photo, the resolution and type of the displaying screen, the illumination conditions of the scene, and the characteristics of the system's camera, may also have a significant effect on the performance of any anti-spoofing algorithm. Moreover, a robust anti-spoofing algorithm should be able to cope with previously unseen attack methods, which were not anticipated prior to its deployment.

Traditionally, face anti-spoofing is approached as a binary classification problem and classifiers are trained on specialised datasets, containing both client and imposter images and videos. The main limitation of this approach is associated with the high cost of creating such databases. That is, a limited only number of attacks is simulated, on a limited number of subjects, while the variability of important environmental factors such as illumination conditions and background is also limited. As a result, the classifiers do not always generalize well to previously unseen attacks. In this Chapter, we aim to answer these questions, the first question is **RQ1**: How to mitigate the problem of imbalanced datasets in face anti-spoofing based on anomaly detection? and the second question is **RQ2**: to which extent the performance of the anomaly detection method will be enhanced if, in the training set, we include images from the in-the-wild and images from non-specialized face databases? In this context, anomaly detection, using classifiers trained on a one-class dataset of client images only, is becoming an increasingly popular approach to face anti-spoofing [27] [35]. The present work is motivated by the observation that training with client images only can also use in-the-wild face images, that is, a set of face images harvested online, as well as face images from databases that do not specialize in face-anti-spoofing.

After giving a brief overview of the general literature on face anti-spoofing in Section 2.5, in Section 3.3 we review the relevant literature on the use of anomaly detection for face anti-spoofing and established our main observation. That is, in the existing literature, the training data are drawn from specialised face anti-spoofing

databases, even though they are just common face images. In Sections 5.3 and 5.4, we describe a proof-of-concept experiment on the feasibility of an alternative approach to the creation of one-class training sets. In particular, we augment an initial training set of client images from specialised face anti-spoofing databases, first with images from non-specialised databases, the SCFace [84] and the CASIA-Web Face [83] in particular, and then with images from the in-the-wild, which were semi-automatically harvested from online sources.

Our anomaly detection anti-spoofing algorithm is based on a Convolutional Autoencoder (CAE). Following a well-established methodology, the CAE is trained on client images, and test images are classified as clients when their reconstruction error is below a threshold. First, we trained the CAE with client images from the Replay-Attack [6] database, and tested it on the Replay-Attack and NUAA [5] databases, creating a baseline. Next, we added into the training dataset the images from the in-the-wild, which were semi-automatically collected from online sources, and finally, we added to the training set images from SCFace and the CASIA-Web Face, which do not specialize in face anti-spoofing. The results show that the classifier's discriminative power, as measured by the Area Under the Curve metric, increases markedly on the unseen NUAA, with a moderate only drop on Replay-Attack. Finally, we added to the training set images from databases that do not specialize in anti-spoofing, SCFace [84] and CASIA-Web face [83] in particular, obtaining again similar results. The main contributions of this Chapter are:

- We review the literature on anomaly detection for face anti-spoofing and establish the observation that the training sets consist of images drawn from specialised face anti-spoofing databases.

- In a proof-of-concept experiment, we developed an anomaly detection method for face anti-spoofing based on a convolutional autoencoder and tested it on the previously unseen NUAA database, showing performance increases when we add into the training set in-the-wild face images and face images from non-specialized databases.

The rest of the Chapter is organised as follows. In Section 5.2, we review the

related work. In Section 5.3, we explain the proposed methods in this Chapter. In Section 5.4, we explain in detail the experimental setup, including data collection. In Section 5.5, we present the results and discuss these results. Then, we briefly conclude in Section 5.6.

5.2 Related work

5.2.1 Autoencoders for Face Anti-spoofing

Autoencoders are a type of neural network that can learn to reconstruct input data, capturing their essential features in a lower-dimensional latent space. This property makes them useful for detecting anomalies or outliers in data. Chen et al. [142] used a pretrained-model (U-Net) based on an autoencoder to reconstruct the original images for the live samples, and zero maps for the spoof ones. The network was guided towards improved reconstruction outcomes using the SSIM and L1 loss functions. In Chapter3, the papers [121,123,143] utilized autoencoder architectures in the context of face anti-spoofing with anomaly detection. The use of autoencoders in face anti-spoofing leverages their ability to learn a compact representation of facial features and their capability to detect deviations from the learned patterns. By comparing the reconstructed image with the original input, the model can identify discrepancies caused by spoofing techniques such as using printed photos or replayed videos as presentation attack instruments. These papers demonstrate the effectiveness of the autoencoder-based approaches in face anti-spoofing for anomaly detection.

5.3 Methods

Our proof-of-concept experiment is based on a convolutional autoencoder (CAE) which is the improved vision of a vanilla autoencoder. Autoencoders are neural networks consisting of two parts. First, the *encoder* processes the input image and produces the *code*, a compressed representation of the input which, usually, has a much lower dimension. Then, the *decoder* reconstructs the original image from the code. The encoder network in a CAE is typically composed of a series of convo-

lutional layers and pooling layers that reduce the spatial dimensions of the input image while increasing the number of feature maps. The decoder network then takes the compressed representation and upscales it back to the original image size using a series of transposed convolutional layers. The goal of a CAE is to learn a compact representation of the input image that can be used to reconstruct the original image accurately. During training, the CAE minimizes the reconstruction loss, which measures the difference between the input and the reconstructed image. This forces the CAE to learn useful features and representations of the input image that can be used for reconstruction. Convolution Autoencoders work better than traditional autoencoders because of their ability to better retain the spatial information of the input image during encoding and then extract this information gradually by what is called the Convolution layers. CAEs have been used in various applications, including image denoising, anomaly detection, and generative models. They can also be used for feature extraction, where the encoder network is used to extract features from the input image. In contrast, at inference, the work of the decoder network is discarded, an approach we employed in some of our experiments in Chapter 7.

The loss function is the reconstruction error, here the Mean Squared Error (MSE) between the original Y and the reconstructed image $\hat{Y}$, seen as $3mn$-dimensional vectors, i.e., one dimension per pixel, per colour band.

The provided formula 5.1 represents the Mean Squared Error (MSE), a common metric used to quantify the average squared difference between actual data points and their predicted counterparts.

In the formula, 'Y' denotes the actual values, and 'Ŷ' represents the predicted values (estimated) value of the data point at position (i, j) for the k-th channel. The MSE calculation involves summing the squared differences for all data points within the dataset, making it a valuable tool for assessing the accuracy of predictions or estimations by measuring the overall error between actual and predicted values.

$$\text{MSE} = \frac{1}{3mn} \sum_{k=1}^{3} \sum_{j=1}^{m} \sum_{i=1}^{n} \left(Y_{i,j,k} - \hat{Y}_{i,j,k} \right)^2 \tag{5.1}$$

As the network is trained to minimise the reconstruction error of the client images,

a high reconstruction error indicates images outside the client class. Thus, we classify an image as an imposter when the reconstruction error exceeds a predefined threshold as in Figure 5.2. Following [144], we implemented both the encoder and the decoder as multi-layer CNNs. Also, following [123], the autoencoder was trained with images from a single class, here are the client images.

5.3.1 Anomaly scores

A straightforward approach to identifying whether an image is imposter involves computing its anomaly score based on how accurately it is reconstructed. The assumption here is that the decoder network will reconstruct client images more effectively as it is trained on them. In contrast, when an imposter (abnormal) image is reconstructed, the outcome will likely be inferior because the network was not trained to handle such images. This thesis uses the MSE as the anomaly score based on the reconstruction error. The reconstruction-based score is the Mean Squared Error (MSE), which is a pixel-level loss that assumes pixels are independent of each other.

Figure 5.1: Example of the real and reconstructed image with their MSE for both client and imposter classes.

Below are examples of the reconstructed images for both clients and imposters, clearly illustrating that the anomaly scores for the imposters are consistently higher than those for the clients. Additionally, this observation indicates that the reconstructed imposter images generally have lower quality.

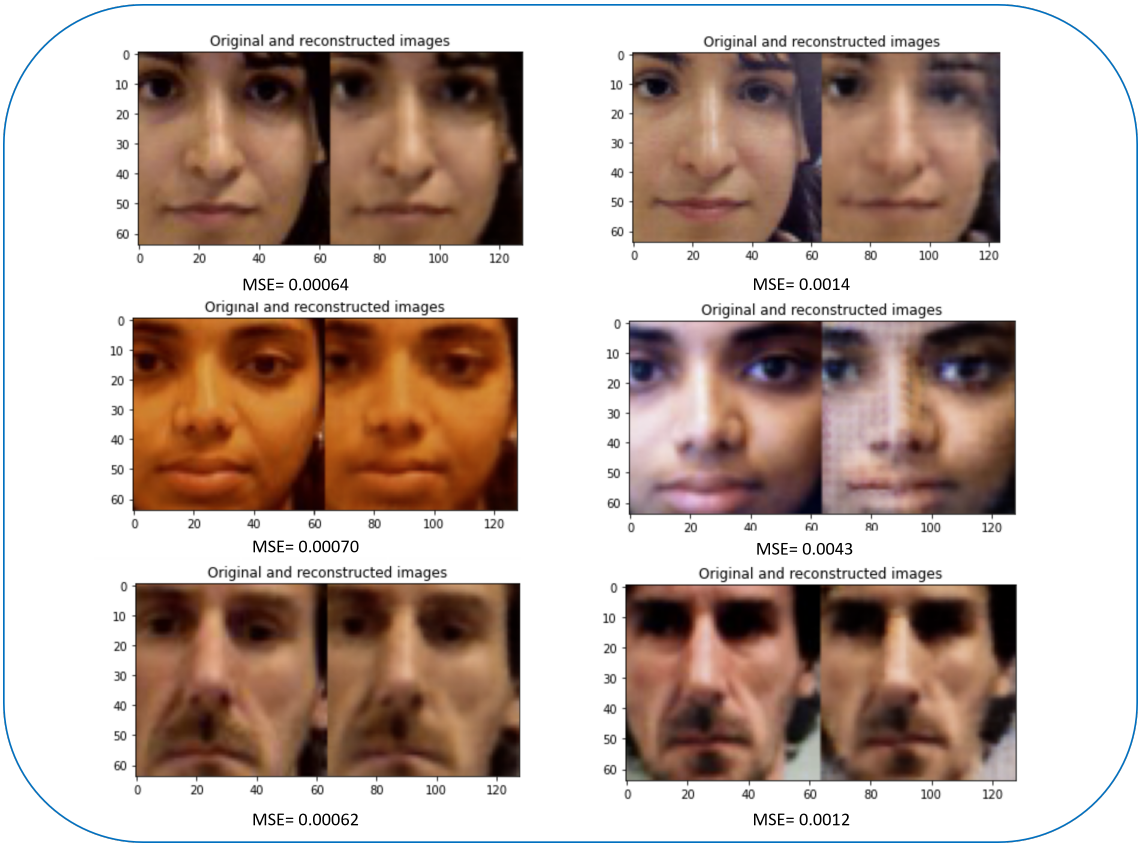

Figure 5.2: On the left, client images alongside their corresponding reconstructed images generated by the CAE, along with the associated MSE values. On the right, impostor images are presented alongside their reconstructed counterparts, also accompanied by their respective MSE values.

5.4 Experiment

5.4.1 Convolutional Autoencoder

Figure 5.3 illustrates the proposed architecture of the autoencoder used in this study. The input is a 64×64 RGB image; the encoder consists of three convolution layers of kernel size (3,3), each one followed by a MaxPooling layer of kernel size (2,2), which is used for spatial down-sampling. The decoder consists of two transpose convolutional layers, followed by one convolution layer, and reconstructs a representation of the original input image. In all layers, we used ReLu activation functions, except for the last layer where a sigmoid function was used.

All code was written in Python, on the Keras platform, and the experiments ran on an Intel Core i7 CPU 64 GB RAM PC with an Nvidia GTX 1650. The whole

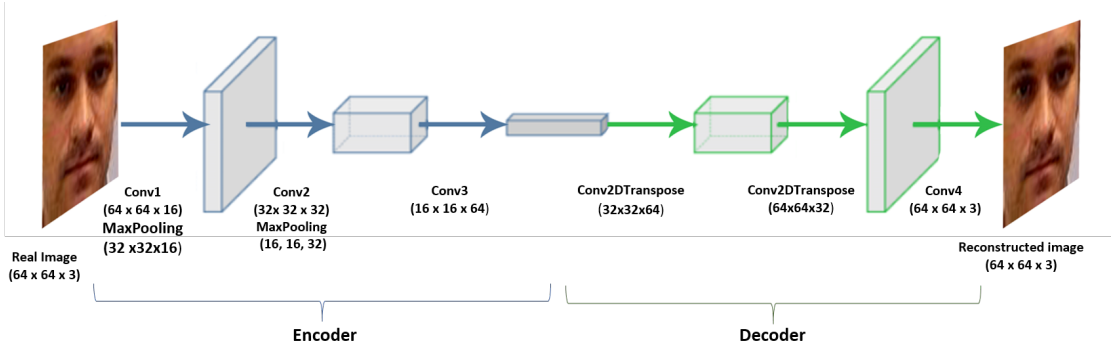

Figure 5.3: The architecture of the proposed convolutional autoencoder.

network was trained with the RMSprop optimizer for 50 epochs, with a learning rate of 0.001. The batch size was set to 32.

5.4.2 Training, validation, and test datasets

The faces in all training and testing images were detected with the Haar feature-based cascade classifier [87], followed by manual inspection and selection. The user input was necessary, especially in the creation of training images from the in-the-wild, due to performance issues of the face detector on such images; general image quality issues such as out-of-focus blurry faces; and in few cases by the need to exclude imposter images, e.g. faces on a poster on a wall. All selected face images were cropped and normalized to 64×64 pixels. We note that face detection followed by cropping is a standard procedure in PAD. In fact, as a standard practice, the images in the benchmark databases are accompanied by a set of coordinates giving the positions of the faces.

The progress in the field of face anti-spoofing is inextricably linked to the development of specialised, image and video, anti-spoofing databases. As we mentioned earlier in Chapter 2, the NUAA Photograph Imposter Dataset [5] is the first such database to become publicly available. The NUAA samples were collected from 15 subjects, using a cheap webcam, in three sessions in different environments and illumination conditions. The attacks were based on digital images captured with a professional camera, and then printed on paper at various resolutions. One commonly used and notable database in the literature is the Replay-Attack [6] database. In our experiment, we will use this database as a baseline. The Replay-Attack database

contains a variety of video presentation attacks, a sufficient number of subjects, different types of simulated attacks, and variability in environmental conditions. These characteristics make it a valuable resource for studying face presentation attack detection.

We tested the autoencoder on two test datasets, the first from the Replay-Attack and the other from the NUAA, consisting of 236 images each. The imposter subset contained images from all types of attacks supported by these two databases. We note that NUAA is considered a particularly challenging case when one is testing for cross-database generalization [145], and the use of webcams as capture devices increases further the challenge of generalisability.

The aim of this Chapter is not to propose an optimised network architecture, but we focus instead on studying the effect of the training set on generalisability. Thus, the architecture and the training protocol of the autoencoder are fixed, and the main variable of our experiment is the training set. To see how the augmentation of the training set with images from non-specialised databases affects the generalisation power of the classifier across the two test sets, we opted for three training sets such that D1 is a subset of D2 and D3, and D2 subset of D3:

D1. Images form the Replay-Attack dataset only. We used 10 client subjects' videos, both controlled and adverse.

D2. We added to D1 102 face images harvested online using general keywords such as *teachers*. These 102 face images were manually chosen from a larger collection, the main considerations being to be frontal face images, in-focus, and of a good size so the normalisation to size 64×64 does not require excessive zooming.

D3. We added 520 images from the SCFace [84] and the CASIA-WebFace databases [83]. The SCFace is a surveillance camera face database from which we used the mugshot, still color images captured indoors under controlled illumination conditions. The CASIA-WebFace is a very large dataset, consisting of 10,575 subjects, collected in a semi-automatic way from the Internet. We used a random subset of it.

Table 5.1 summarizes the description of the training datasets. Also, in Figure5.4 some examples of the wild images and images from non-specialised DBs.

Table 5.1: Description and size of the training datasets.

	Description	**size**
D1	Replay-Attack	2800
D2	Replay-Attack + in-the-wild mages	2902
D3	Combine Replay-Attack with others DB	3422

The validation dataset was kept constant to simplify the design of the experiment. It consisted of 578 live and fake images from the Replay Attack. As the use of a validation set with a composition similar to the most general training dataset **D3** may lead to an underestimation of the performance of the proposed autoencoder on the Replay-Attack under an intra-database protocol, we also report HTER values computed under the use of a validation set consisting of Replay-Attack images only.

Figure 5.4: The first row are Examples of wild images and the second row are examples from face recognition DBs.

5.5 Results and Discussion

Figure 5.5 shows the ROC curves of the proposed autoencoder, trained on the three datasets, and tested on Replay-Attack (up) and NUAA (down). The corresponding Areas Under the Curve(AUC) are reported in Table 2. We notice that the inclusion of the in-the-wild images in the training dataset improved markedly the cross-database generalisation power of the classifier, with the value of the AUC on the NUAA going up from 0.63 when trained with **D1** to 0.72 when trained with **D2**. Moreover, the inclusion of images from non-specialized databases further increased the AUC to 0.80 when trained with **D3**. There is also a noticeable fall in the performance on Replay-Attack, with the AUC going down from .93 to .89 and then to .82. We also note the high performance of the algorithm under an intra-database test mode, that is, the high AUC value of .93 AUC on the Replay-Attack.

Table 5.2: The AUC values corresponding to the ROC curves shown in Figure 5.5.

	D1	**D2**	**D3**
Replay-Attack	.93	.89	.82
NUAA	.63	.72	.80

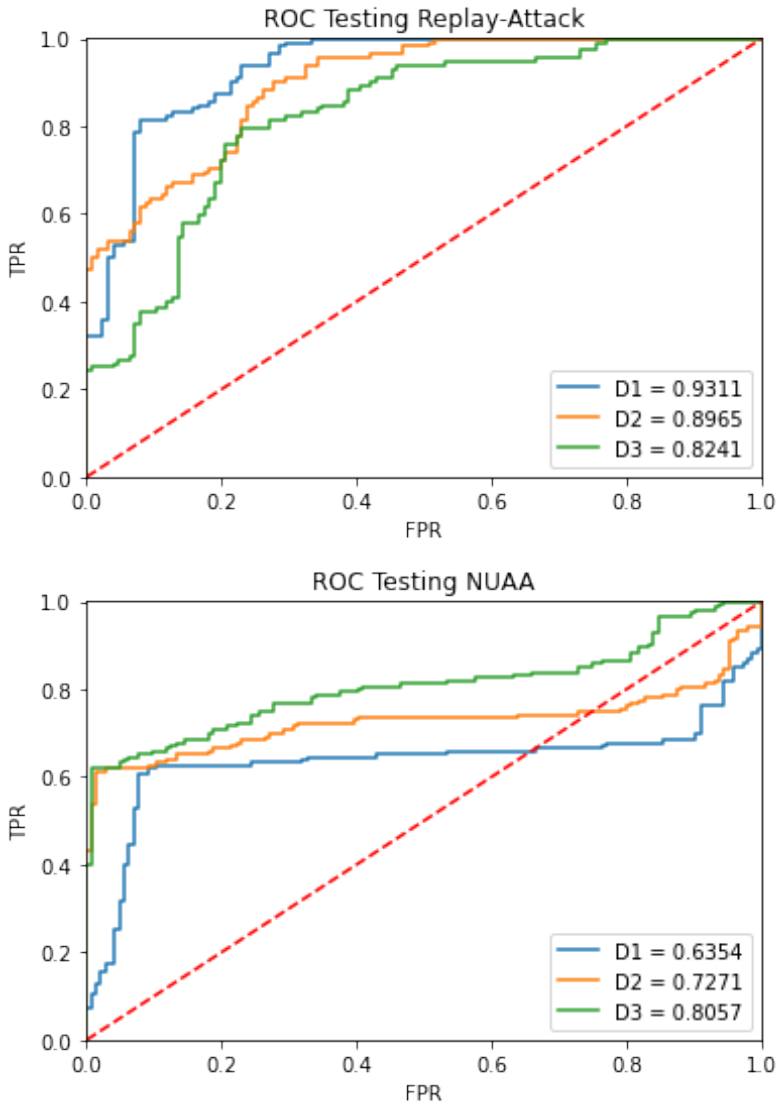

Figure 5.5: ROC curves corresponding to classifier/training dataset combinations, tested on Replay-Attack (top) and NUAA (bottom).

The value of the AUC is an integral over all possible operating points, that is, overall possible thresholds against which we compare the reconstruction error to determine whether a sample should be classified as client or imposter. Thus, it separates the problem of assessing the discriminative power of the classifier from

the problem of finding an optimal, for the given test, operating point. Next, we will discuss the problem of determining an optimal operating point.

In the literature, classifier performance on a specific operating point is usually assessed either by reporting separately the False Positive Rate (FPR) and the False Negative Rate (FNR), or their mean average Half Total Error Rate. We note that reporting an operating specific performance metric does not necessarily mean that the problem of finding the optimal operating point has been addressed. For example, some papers report the minimum HTER overall operating points or the True Positive Rate corresponding to certain fixed values of FPR. Employing a technique that is commonly used to address this problem, we first compute a threshold on the validation set, here the threshold corresponding to the Equal Error Rate (EER) on that set, and use this threshold to compute the HTER.

Table 3 summarizes the HTERs of our method, and for comparison, the HTER range of the four different classifiers proposed in [11]. Regarding the performance on the NUAA, despite the satisfactory discriminating power of the classifier as shown by the ROC curves, the high HTER values indicate that the threshold computed on the validation set, which is a set containing images from the Replay-Attack only, cannot be used on the NUAA. We note that [11] also reports very high HTERs, which again indicates that a satisfactory operating point on NUAA could not be found. Regarding the performance on the Replay-Attack, we note that under such an intra-database testing mode, our convolutional autoencoder performs significantly worse on Replay-Attack than the best performing classifier in [11], but its performance is inside their reported range.

Table 5.3: HTERs computed on the operating point corresponding to the EER of the validation set. The last column shows the range of HTERs reported in [11].

	D1	**D2**	**D3**	[11]
Replay-Attack	.15	.24	.26	[.05 - .32]
NUAA	.57	.54	.39	[.51 - .65]

5.6 Conclusions

Research of the literature on anomaly detection methods for face anti-spoofing showed that their training datasets are exclusively drawn from specialised spoofing databases. However, a one-class training set for anomaly detection methods does not require specialized images; these can be obtained from non-specialised databases or from the wild. In a proof-of-concept experiment, we showed that the inclusion of such images in the training set of a convolutional autoencoder, which was originally trained on the Replay-Attack database, increased its performance on the unseen NUAA database, as shown by the ROC curves and the corresponding AUC values while, conversely, the performance on the Replay-Attack itself decreased. That is, as expected, the inclusion of images from outside the specialised databases had a moderating effect, rather than an always positive or an always negative one. In a cross-database testing mode, performance increased, in an intra-database testing mode performance decreased. We note that in the most recent papers on face anti-spoofing, as the ones reviewed in Chapters 2 and 3, cross-database testing is becoming the norm. That is, regardless of its relevance in typical practical application scenarios, in the literature, the issue of how the classifier would perform on images from previously unseen databases is considered important. In this context, the behaviour of the classifier on unseen client images outside specialised anti-spoofing databases becomes an equally legitimate question. However, there are various methodological challenges arising from the use of such test sets, since the client images will be drawn from much more diverse sources than the imposter images. This is an issue we plan to address in our future work.

CHAPTER 6

Anoformer PAD: Anomaly Detection with Transformers for Face Anti-spoofing

In Chapter 5, our research demonstrated that anomaly detection methods that solely rely on training with bona fide images can outperform traditional binary classification approaches in a cross-database testing mode, indicating superior generalisation properties. This finding is significant because, in real-world scenarios, the system might be unaware of any information about imposters. Based on this insight, here we propose a novel model that only extracts deep features from bona fide images, eliminating the need to incorporate impostor images during the training phase. To achieve this, we leveraged the power of transformers, which in many settings have been proven effective in extracting these essential features. To the best of our knowledge, there is no study of transformer-based anomaly detection techniques for PAD.

6.1 Introduction

While face recognition is the biometric authentication method of choice in many application domains, it is still considered extremely vulnerable to presentation attacks. In such attacks, an imposter is trying to gain unlawful access by presenting in front

of the system's camera a printed photo, or an electronic screen playing a video of a rightfully registered person. The vulnerability of face recognition systems to such spoofing attacks means that they cannot be safely deployed in security-sensitive applications in uncontrolled environments, as for example ATM machines in the high street. Presentation attack detection (PAD) addresses this problem by developing binary classification algorithms aiming at distinguishing between the genuine, bona fide samples presented to the system's camera, and the imposter ones.

The most common approach to PAD is to train a binary classifier on both the bona fide and the imposter classes. In this case, training and testing are performed within specialised face anti-spoofing databases, which due to the high cost of producing imposter samples have limited variability, raising questions on the generalisation power of the classifier, especially on unseen attacks in scenarios that have not been covered by the testing databases. In particular, while the current state-of-the-art algorithms can show good results on unseen attacks within the same database, and some generalisation power between specific databases, a thorough cross-database validation is expected to show that they do not always generalise well. For example, in [62] all the eleven methods under comparison show HTERs between 24% and 60.6% in cross-database generalisation task from the Replay Attack database to the CASIA-MFSD.

An alternative approach aiming at addressing the generalisation problem is anomaly detection based on one-class training. We note that in the limited testing environments provided by the existing databases, anomaly detection approaches underperform two-class training under most testing protocols. However, they have the conceptually appealing property that they neither attempt to learn specific presentation attacks nor, most importantly, specific environments where such attacks were modelled during the creation of the database. Thus, anomaly detection for face anti-spoofing is still a very active research area [27, 33].

This Chapter concentrates to answer the fifth research question which is **RQ5**: Can a deep generative model perform Anomaly Detection efficiently and solve the generalization problem? Therefore, we proposed to use the Vision Transformer (ViT) [10] and the ResNet [125] as backbones for anomaly detection for face anti-

spoofing. Our motivation for using ViT was the observation that in several computer vision tasks Transformers are replacing Convolutional Neural Networks (CNNs) as the new gold standard. They have already been proposed for the PAD problem under a two-class training setting [146], but they have not been used yet for PAD in the anomaly detection setting.

Regarding the use of ResNet, we note that the size of the receptive field is one of the primary distinctions between a CNN-based model and a transformer-based model. Whereas due to the self-attention mechanism, the transformer is superior in its ability to capture a pixel relation over a long distance [147]; nonetheless, it lacks a reliable way of capturing spatial information within each patch so it may overlook crucial spatial local patterns, such as textures. However, CNNs are different in this regard, focusing on textures rather than shapes to identify objects in images [148]. ResNet, in particular, is a highly efficient neural network architecture, and its residual learning methodology addresses the degradation issue which exists in many other CNN models. Thus, overall, we leverage the strengths of two state-of-the-art architectures, a transformer and a CNN to extract reliable features. Our ablation study shows that the combined ViT ResNet backbone gives significant improvement over a single network backbone.

Our main contributions are summarised as follows:

- A novel Anomaly detection Vision Transformer (AnoFormer), with ViT and ResNet in the backbone, for presentation attack detection.
- A comparison of various one-class classification models, showing that a decoder with MSE as a loss function outperforms the other configurations.
- An ablation study showing that the use of a combination of ViT and ResNet in the backbone outperforms the use of single networks.

The rest of the Chapter is organised as follows. In Section 6.2, we review the related work. In Section 6.3, we explain the proposed methods in this Chapter. In Section 6.4, we present the proposed AnoFormer and its implementation details. In Section 6.5, we present the results, and we briefly conclude in Section 6.6.

6.2 Related Work

In this section, we focus on discussing the related work that specifically pertains to the topic of Transformer in face anti-spoofing. While Chapter 2 provided a general overview of the broader field of face anti-spoofing, we now narrow our focus to studies and research specifically related to Transformers in this domain.

6.2.1 ViT Transformers for face anti-spoofing

Inspired by the Transformer scaling successes in Natural Language Processing (NLP), authors in [10] applied the standard Transformer directly to images with image patches as an input to the Transformer with a sequence of linear embedding of these patches. Image patches are treated in a manner analogous to tokens (words) in NLP applications, and the model is trained using supervised image classification techniques.

Although numerous methods have been suggested for detecting spoofing attacks, many of them suffer from overfitting to the training set and fail in generalizing to unseen attacks and environments. In a study conducted by George et al. [146], they explored the efficacy of the vision transformer model for addressing the zero-shot presentation attack detection problem. The proposed method was evaluated alongside both state-of-the-art methods and fine-tuned CNN models. They investigated this method's performance under challenging conditions, including unseen attack scenarios and cross-database scenarios.

Wang et al. use their proposed Transformer-based face anti-spoofing model in their work [149]. Authors invoke a sequence of multi-transformer encoder layers to achieve more comprehensive local patches for two novel models; the use of cross-layer relation-aware attentions (CRA) and hierarchical feature fusion (HFF). Huang et al. [150] use an adaptive transformer model to handle challenging print and replay spoof attacks across different datasets using a few-shot setting. As mentioned in the chapter's introduction, to the best of our knowledge, there is no study of transformer-based anomaly detection techniques for PAD.

6.3 Methods

In this section, we will provide a concise overview of the background of the methods used in this chapter. The aim is to briefly explain the underlying concepts and principles behind these methods.

6.3.1 Transformer

The Vision Transformer (ViT) is a deep learning model architecture with an impressive performance on various computer vision benchmarks. In particular, it is competitive against and sometimes surpasses traditional convolutional neural networks in image classification tasks. Additionally, ViT has been successfully applied to tasks like object detection and image segmentation.

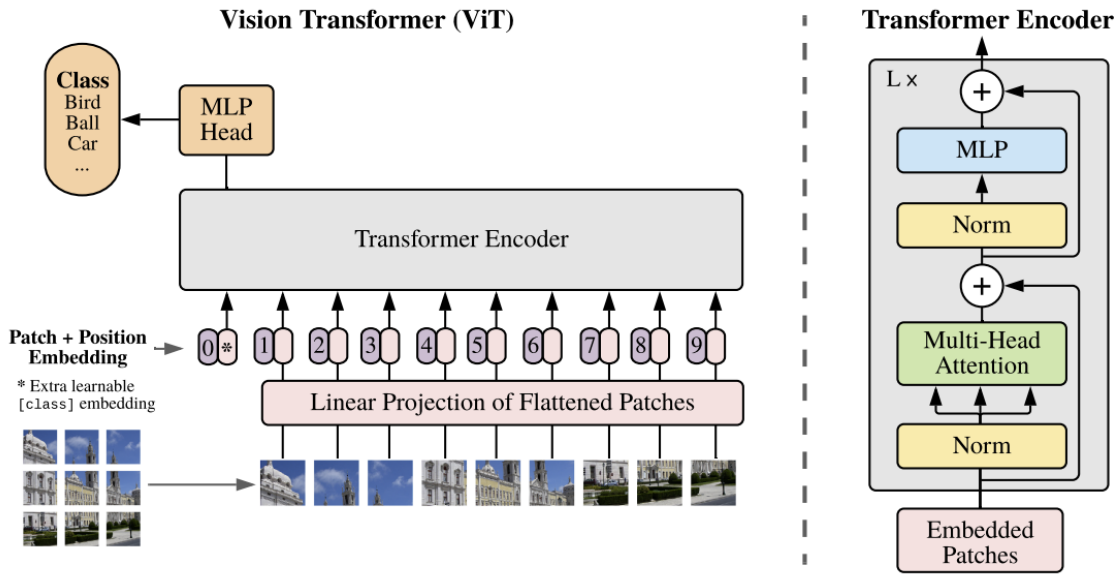

Figure 6.1: Architecture of the Vision transformer from [10].

Figure 6.1 provides an overview of the essential components of the vision transformer. To process an input image, we divide it into patches of equal size. Each patch represents a local region of the image. These patches are then flattened into vector representations, called patch embeddings, which serve as input tokens for the Transformer model. To capture spatial information and preserve the positional relationships between patches, positional encoding is added to the patch embeddings.

This allows the model to understand the layout of the patches within the image. In addition to the patch embeddings, a unique CLS token is added at the beginning of the patch sequence. The CLS token stands for classification tasks which represents the entire image and carries global information. The Transformer model uses this CLS token to capture high-level image features and make predictions based on them.

The transformer encoder layer, originally proposed in [10], consists of multiple encoder blocks. Each block incorporates multi-head self-attention (MSA) and multi-layer perceptron (MLP) components. Self-attention enables each patch embedding to attend to other patch embeddings, capturing relationships and dependencies. The MLP processes the encoded patch embeddings independently, applying non-linear transformations to capture complex patterns and interactions. Layer normalization (LN) is applied after each self-attention and feed-forward network layer, helping to stabilize the training process and improve the model's generalization.

Finally, the output layer takes the final encoded representations and performs classification tasks, such as predicting image class labels. It typically consists of fully connected layers and softmax activation for multi-class classification. Combining these components, the ViT model transforms image classification tasks into sequence-to-sequence problems, treating the patches as tokens. This approach allows the model to leverage the Transformer architecture's power in capturing global and local information in images.

6.3.2 Frechet Inception Distance (FID)

The Fréchet Inception Distance (FID) [151] is a metric commonly used to assess the quality of generated images compared to real images. We utilise the FID as an anomaly score in this Chapter. The FID is employed as a measure of dissimilarity between the generated and genuine samples. A higher FID score indicates a more significant deviation from the genuine distribution, indicating the presence of anomalies. Therefore, the FID score serves as an anomaly score, helping to identify and classify abnormal instances in the dataset [152].

6.3.3 One Class Classifiers

One Class SVM

One-Class SVM (Support Vector Machine) [100] is a machine learning algorithm for anomaly detection. Unlike traditional SVMs used for binary classification, One-Class SVM is trained on only one class of data, considered the normal or target class. It aims to identify instances that deviate significantly from this normal class. The main idea behind One-Class SVM is to find a hyperplane that encloses most of the target class instances, effectively creating a boundary separating the normal samples from the outliers. This hyperplane is constructed to maximise the margin between the hyperplane and the target class instances. The instances that lie outside the margin are considered anomalies or novelties.

Isolation forest

Isolation Forest [153] is an unsupervised machine-learning algorithm utilised for anomaly detection. It is based on isolating anomalies rather than identifying normal instances. The algorithm constructs a forest of random decision trees and isolates anomalies by creating short paths in the tree structure. The main idea behind Isolation Forest is that anomalies are often less frequent and more easily separable from normal instances. By randomly selecting features and splitting values, the algorithm can quickly isolate anomalies, as they require fewer splits to be separated from the rest of the data.

6.4 The Anoformer

The proposed method uses feature vectors provided by the pre-trained ViT and ResNet [125], which are then processed by a one-class classification technique. In our experimental study in Section 6.5 we show results obtained by the use of isolation forests and one-class SVMs. However, our focus is on training a decoder of the feature vectors and then comparing the reconstruction error against a threshold to take the classification decision.

6.4.1 Architecture

The architecture of the Anoformer is illustrated in Fig. 6.2. The backbone networks are ViT, which, has already demonstrated its potential as an embedding extractor for the face PAD problem in [146] where a two-class training of the ViT feature vectors gave results competitive to the state-of-the-art, and ResNet-18, both pre-trained on ImageNet [154].

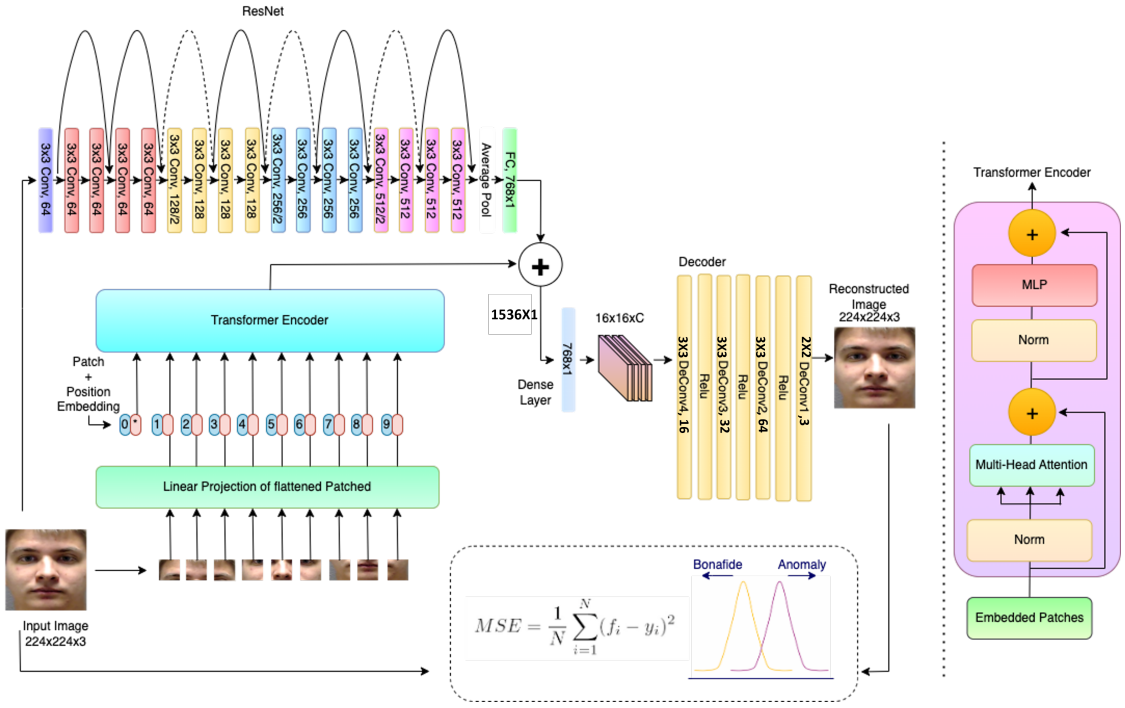

Figure 6.2: Architecture of the Anoformer.

Regarding the choice of specific ViT architecture, we first note that our proposed model can work with any version of ViT, providing compatibility with future improvements to ViT. Here, we employed the Data-Efficient Image Transformer [155] (DeiT-Base), which is an improved version of ViT of lightweight design. To learn diverse features, the training dataset's bona fide images are fed into the ViT and ResNet networks. The ViT divides the input image into patches and uses the extracted features as the sequence input for the transformer, followed by transformer layers. The transformer encoder layer is composed of multiple encoder blocks, each with multi-head self-attention (MSA) and multi-layer perceptron (MLP), as in [10].

In Equation 6.1, we encounter the "Attention Mechanism," a crucial concept in AI. It's like a spotlight that decides which data points are more important by

comparing queries (Q) with keys (K) and using Softmax. This decision guides the model's understanding of the data, making it proficient in tasks like language comprehension and image analysis. In practical terms, this mechanism involves transforming image patches into queries, keys, and values, along with position encoding (PE) added to keep track of each input token's position. The MLP contains two linear layers with a GELU (Gaussian Error Linear Unit) activation function. Finally, the encoded patches are reshaped and projected into a reconstruction vector via a learned projection matrix.

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{Q \cdot K^T}{\sqrt{d_k}}\right) V \tag{6.1}$$

The ViT leverages the attention mechanism which gathers information from the entire input sequence. Self-attention layers scan through a sequence of elements and update them based on the information obtained from the entire sequence. In essence, they simulate explicitly every pair-wise interaction that occurs between the components of the input sequence. Thus, the self-attention maps, which are learned separately for each layer, are necessary for a transformer model to encode the dependencies between input tokens.

Fig. 6.3 shows attention maps from different ViT layers for a bona fide and an imposter input. We note the difference in ViT's behaviour between the two classes, in that certain prominent facial features such as eyes, nose, and mouth are more prominent in the imposter attention maps.

The decoder decodes the 768×1 reconstruction vector back to the original image shape. We used four transposed convolutional layers, with ReLU in between, except for the last layer, which is followed by a sigmoid as the final activation function. The decoder part of the Anoformer was trained with the features of the bona fide images extracted from ViT and ResNet. The decoder is trained with the objective of minimizing the error between the input and the output of the network, aiming at reconstructing bona fide images with high fidelity.

6.4.2 Implementation

To avoid contributions from the input image's background, we use the MTCNN algorithm for face detection, and cropped the images, retaining the face regions only. Then the images are rotated to have the eye centres horizontally aligned and finally resized to 224×224, which is the native resolution of the ViT transformer pre-trained on ImageNet.

In the final binary classification section, we used an Adam optimizer with an initial learning rate of 1e-4 and batch size 16. A label-smoothing cross-entropy loss function was used to train the classifier. The MLP head is the binary classifier which contains two fully-connected layers of dimensions 512 and 2. The development environment was the PyTorch running on a PC with an Intel CPU, 64 GB of RAM, and Google Colab GPU.

Our backbone consists of two parts, the first part is the Deit-Base [155], which has spatial position embeddings to improve image processing capabilities and has an output embedding dimension of 768. The second component is a ResNet-18 network with 18 layers that have once more been trained on ImageNet. The size of the output from the ResNet is 2048. Therefore, in order to bring the size of the outputs (features) produced by ResNet down to the same level as those produced by transformer 768, we add one dense layer at the very end of the network. Finally, then concatenate the two feature sets, the one from Deit-Base and that from Resnet-18, to create a vector of 1536 features. This data is then compressed to a size of 768 by adding one dense layer before being sent to the decoder.

The decoder was trained under the Mean Squared Error (MSE) loss function. While it is a pixel-level loss, assuming independence between pixels, it has been repeatedly shown that it works very well in practice, and its simplicity and the fact that it was supported by our development environment led to fast training times.

We used random search for fine-tuning the model, which is computationally less expensive than grid search because it doesn't explore every possible combination. However, it can still effectively find good hyperparameters because it samples a diverse range of hyperparameter configurations. This diversity often leads to quicker convergence and can be particularly useful when computational resources are lim-

ited. Random search is a popular choice in hyperparameter optimization and is widely used in practice for tuning machine-learning models.

Regarding the computation of the anomaly score, that is, the error between the input and the reconstructed image, the natural choice is to use the loss function itself. Thus, we use MSE as our default. We experimented with other error metrics, such as Cosine Similarity, the Structural Similarity Index (SSIM), and the Frechet Inception Distance (FID) score [151]. We found that FID performed comparably to MSE, even though the decoder was trained with MSE, and thus, we include some relevant results in Section 6.5.

6.5 Results and Discussion

6.5.1 Databases

Our experiments were performed on two commonly used face anti-spoofing databases, the *Replay-Attack* (RA) [156], and the *Spoof in the Wild* (SiW) [124].

We divided the RA and SiW databases into training, validation, and testing sets with non-overlapping subjects. In the validation and testing datasets of the RA database, we included all three presentation attack species; printed photo(a1), video (a2), and digital photo (a3). In the validation datasets of the SiW database, we included three representative attack species; printed photo (A1), replay attack using iPhone (A2), and replay attack using a tablet (A3).

6.5.2 Evaluation Metrics

We report our results using the APCER, BPCER and ACER metrics, which are the most commonly used error metrics in face anti-spoofing, recommended by the ISO/IEC 30107-3:2023 [157] protocol for testing and reporting on biometric PAD.

As we mentioned earlier in Chapter 2, the definition of APCER as the maximum of the misclassification rates that are computed separately over each attack species brings to the fore an important methodological problem. When we measure the misclassification rates, do we use a single threshold for all attack species, or do we

choose a different threshold for each one of them?

For example, in [146] different thresholds were used, and thus different BPCERs are reported as corresponding to each attack species, even though the dataset of bona fide presentations is one. Here, we use a single threshold for all attacks, firstly because in practice it is unrealistic to expect prior knowledge of the attack species that will inform the choice of threshold, and secondly, because we think it is closer to the spirit of the ISO definition of APCER, that is, to consider the worst case outcome over all attack species, rather than splitting the problem into smaller, easier to tackle sub-problems. In particular, we used the threshold corresponding to the Equal Error Rate (EER) on an independent validation set from the same database as the testing set.

6.5.3 Anoformer validation

In Tables 6.1 and 6.2 we report the results for four different classifiers over the ViT+ResNet backbone, tested on the RA and SiW databases, respectively. The one-class SVM (OC-SVM) is a widely used one-class classification method, being essentially an SVM trained with positively labelled data only, and aiming at maximising the separation of their class from the origin of the coordinate system. The second classifier we used is the Isolation Forest, which is based on decision trees and it is theoretically justified under the assumption that anomalies are "few and different". We note that while this is a very realistic assumption for the face anti-spoofing problem, it is not reflected in the usual PAD evaluation protocols that we also use here. Finally, in the last two rows of the tables, we report error rates for the Anoformer and the Anoformer with the FID metric for the computation of the anomaly score. We notice that the combination of Anoformer with MSE in the reconstruction gives lower ACERs on both databases, and it is the configuration that we will evaluate.

Table 6.1: ViT + ResNet backbone with various one-class classifiers tested on RA.

OCC	**ACER**	**APCER**	**BPCER**
OC-SVM	.31	.26	.36
Isolation Forest	.33	.27	.40
Anoformer MSE	**.13**	**.23**	**.03**
Anoformer FID	.19	**.23**	.16

Table 6.2: ViT + ResNet backbone with various one-class classifiers tested on SiW.

OCC	**ACER**	**APCER**	**BPCER**
OC-SVM	.31	.16	.46
Isolation Forest	.33	**.11**	.55
Anoformer MSE	**.21**	.33	**.10**
Anoformer FID	.22	.35	**.10**

Table 6.3 shows the results of the ablation study on the backbone of the Anoformer. The ViT + ResNet combination gives on both databases lower ACERs than ViT or ResNet alone, and notably the APCERs and BPCERs are both lower in both cases.

Table 6.3: Ablation study for the Anoformer backbone.

	RA			**SiW**		
	ACER	**BPCER**	**APCER**	**ACER**	**BPCER**	**APCER**
ViT	.16	.07	.25	.38	.42	.34
Res	.19	.07	.31	.44	.36	.52
V+R	**.13**	**.03**	**.23**	**.21**	**.10**	**.33**

6.5.4 Performance evaluation

In Table 6.4, we compare the error rates of the proposed Anoformer against [8], which is a recently published anomaly detection method that reported results on the same databases and with the same error metrics as us. The results show that the Anoformer gives a lower ACER on both RA and SiW.

Table 6.4: Performance comparison against [8].

Database	**Methods**	**ACER**	**APCER**	**BPCER**
RA	[8]	.21	.25	.17
RA	ours	**.13**	**.23**	**.03**
SiW	[8]	.23	**.23**	.23
SiW	ours	**.21**	.33	**.10**

In Table 6.5 we report AUCs of the Anoformer on cross-database generalisation and compare it against a baseline of a two-class training method, where the ViT feature vector is fed to an MLP. We report the AUC of each presentation attack species separately, noting that as AUC is an average performance measure computed over all possible thresholds, the average of these results is an appropriate performance measure over the entire database. We notice that while two-class training significantly outperforms the proposed anomaly detection in intra-database performance,

anomaly detection outperforms two-class training in the RA/SiW cross-database testing, as well as in the SiW/RA cross-database testing in two of the three attack modalities.

In comparison, in RA/RA intra-database testing in table 6.5, the previously proposed anomaly detection method based on live correlation loss [89], achieved AUCs of (.89, .93, .88) on the three RA attack modalities. We note that their second database was OULU and thus, we cannot easily compare our results any further.

Table 6.5: Comparison against ViT feature vector + two-class trained MLP, in intra- and cross-database testing. We report AUC values on each presentation attack species separately.

	ViT(MLP)			**Anoformer(ours)**		
	a1	**a2**	**a3**	**a1**	**a2**	**a3**
RA/RA	.99	.98	.93	.91	.97	.91
SiW/RA	**.98**	.98	.84	.97	**.99**	**.86**
	A1	**A2**	**A3**	**A1**	**A2**	**A3**
RA/SiW	.85	.84	.86	**.93**	**.86**	**.93**
SiW/SiW	.96	.95	.93	.90	.92	.95

In comparison in [158] an overall AUC of 91.2% was reported for the best-performing configuration of LBP + NN.

Finally, in Table 6.6, we report cross-database testing results for the Anoformer with threshold-specific metrics. As expected cross-database testing gives significantly higher ACERs. However, we note that this is mostly due to the higher APCERs. which are affected by a very high error rate in a single attack modality. That is, a .5 in SIW/RA modality 3 (digital photo), and a .42 in RA/SiW modality 2 (replay attack using a tablet). Interestingly, in the other attack modalities, the error in cross-database testing is lower or equal.

Table 6.6: Intra- and cross-database testing of the Anoformer with threshold-specific metrics.

	ACER	**BPCER**	**APCER**
RA/RA	.13	.03	.23
SiW/RA	.26	.03	.50
RA/SiW	.27	.13	.42
SiW/SiW	.21	.10	.33

6.6 Conclusion

This chapter introduced Anoformer, a novel anomaly detection model designed specifically for Presentation Attack Detection (PAD). Anoformer combines the strengths of pre-trained transformers, specifically the Vision Transformer (ViT), and deep Convolutional Neural Networks (CNNs), specifically ResNet, utilising them as the backbone of the model. Additionally, we incorporated a one-class trained convolutional decoder for the reconstruction of the images. The experimental results of our study demonstrate that Anoformer achieves competitive performance within the class of anomaly detection algorithms for generalized face anti-spoofing. Also, our experimental results show that the performance of the model in cross-database testing can outperform a two-class trained baseline. This signifies the effectiveness and potential of Anoformer as an advanced solution for PAD applications.

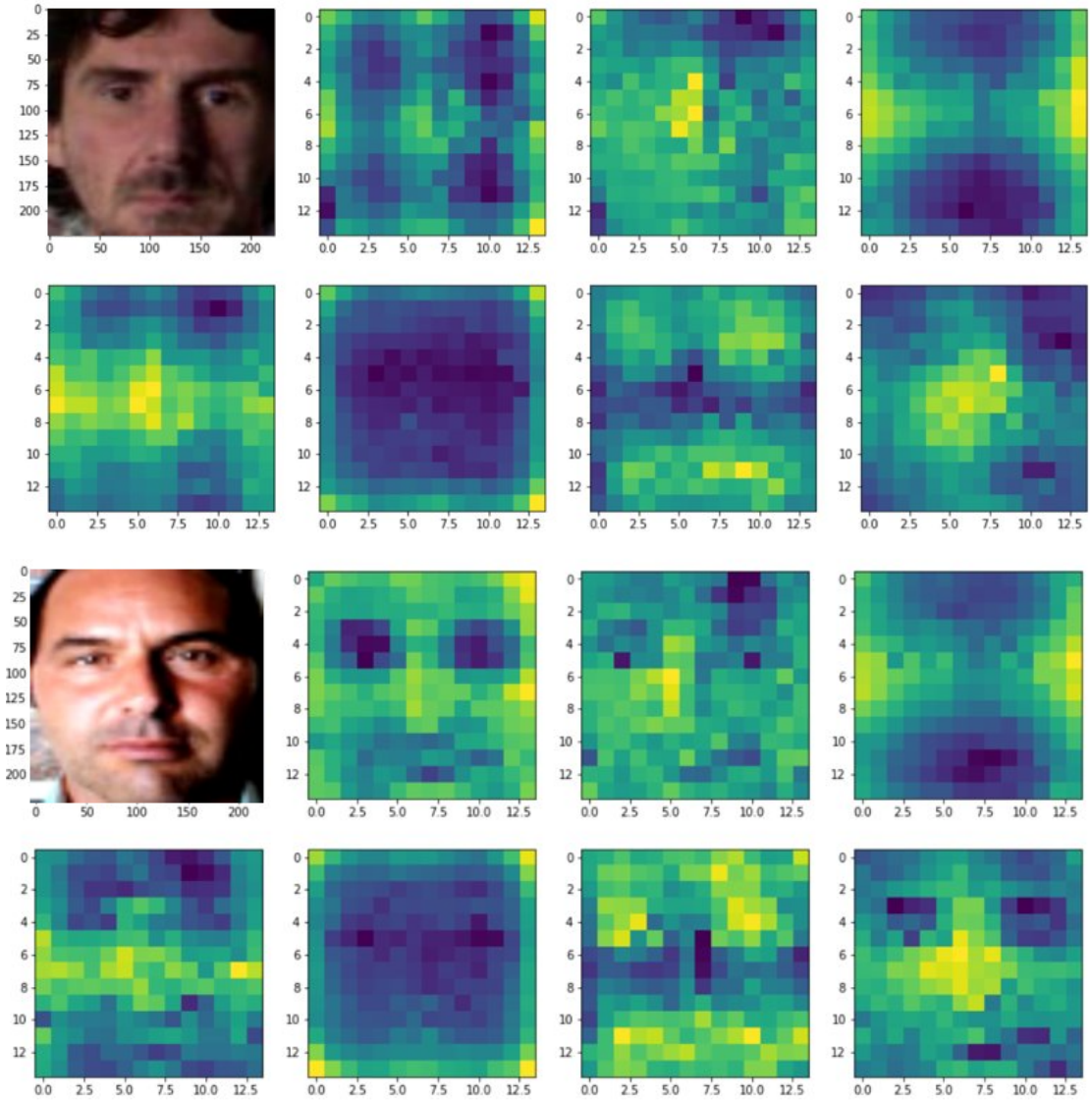

Figure 6.3: Visualisation of input faces from the Replay Attack database, and the attention maps of several ViT layers. **Top two rows:** a bona fide face. **Bottom two rows:** an imposter face.

CHAPTER 7

Race Bias Analysis of Bona Fide Errors in face anti-spoofing

While working on anomaly detection as our method of choice to address the limitations in the area, especially regarding generalisation, we have observed the presence of bias in existing face spoofing databases, particularly regarding racial bias. In the field of face anti-spoofing, researchers are constantly striving to improve models using various techniques. However, a significant portion of research in this area tends to overlook various forms of bias, such as race, age, or gender bias. These biases can significantly impact the performance of anti-spoofing models and raise ethical, legal and regulatory issues regarding their deployment.

Race bias, in particular, is concerned with the potential disparity in the performance of face anti-spoofing models among individuals of different races. Such performance differences can lead to biased outcomes; depending on the application, certain racial groups may experience lower accuracy rates than others. This can result in misidentification, discrimination, or unequal treatment of individuals, especially for those from underrepresented racial groups.

In this Chapter, we delve into an in-depth examination of racial bias, specifically within the domain of face anti-spoofing. Our goal is to shed light on the impact of these biases and contribute to a better understanding of the challenges

and implications associated with racial bias in face anti-spoofing research.

7.1 Introduction

Face recognition is the method of choice behind some of the most widely deployed biometric authentication systems, currently supporting various applications, from passport control at airports to mobile phone or laptop login. A key weakness of the technology, preventing it from being employed in security-sensitive applications in uncontrolled environments, for example, ATM machines for money withdrawal, is its vulnerability to *presentation attacks*, where imposters attempt to gain wrongful access by presenting in front of the system's camera a photo, or a video, or by wearing a mask resembling a registered person. To solve this problem, algorithms for presentation attack detection (PAD) are developed, that is, binary classifiers trained to distinguish between the bona fide samples from live subjects and those from imposters.

The large variety in the types of possible presentation attacks and the large variation in the environmental conditions under which they might take place make PAD a particularly challenging problem. However, the current state-of-the-art, utilising the power of deep learning, comprises classifiers with excellent accuracy rates and a satisfactory generalisation power to at least a limited number of previously unseen attacks. Cross-database generalisation is still problematic. However, it is debatable if this is a real obstacle to deploying PAD algorithms in practical applications since such algorithms are usually embedded in specific face recognition systems with given camera specifications and configurations.

Here, we deal with the problem of racial bias in face anti-spoofing algorithms. It is a topic that has attracted considerably less research interest than accuracy and generalisation power, despite the fact that it raises ethical, legal, and regulatory considerations, which, by their own, can prevent adoption in specific applications. Addressing this gap, the aim of this Chapter is to provide a framework for studying the question: *Does the classifier work equally well on people from all races?*. How can we investigate racial bias in the face anti-spoofing anomaly model?

The proposed race bias analysis process has three key characteristics. First, the focus is on the bona fide error, that is, on genuine people wrongly classified as imposters. Bias in this type of error has significant ethical, legal, and regulatory ramifications, and as it has recently pointed out, "creates customer annoyance and inconvenience, and this is also where bias can occur in PAD systems", [159]. Secondly, we analyse various stages of the classification process. Not just the final binary outcome but also the scalar responses of the network before thresholding and, before that, the representation of the face image in the network's latent space. Thirdly, we treat the threshold value determining the classifier's operating point on the ROC curve as a user-defined variable. We do not assume it is fixed by the vendor of the biometric verification system through a black-box process. In the rest of the Chapter, we demonstrate the application of the proposed bias analysis approach on a face anti-spoofing algorithm based on the recently proposed Vector Quantized Variational Autoencoder (VQ-VAE) architecture, [160]. The network is trained and validated on the Replay Attack and the SiW databases and tested for racial bias on bona fide samples from the SiW and the RFW databases. Hypotheses are tested using the chi-squared test on the binary outcomes, the Mann–Whitney U test on the scalar responses, and Hartigan's Dip [161] for testing bimodality in the response distributions. To test for bias in the latent space of the VQ-VAE network, we train an SVM with encoding vectors from two races and measure its performance as a binary classifier. The experiment presented in this Chapter is based on a one-class trained autoencoder. Anomaly detection is a popular approach [8, 143, 162–164], offering good generalisation to unseen attacks.

The contributions of the Chapter are summarised as follows:

- A demonstration that racial bias can be attributed to several characteristics of the response distributions: different means; different variances; bimodality; outliers.
- A demonstration that non-specialised databases, such as RFW, can be used to analyse face anti-spoofing algorithms.
- A VQ-VAE based network for face anti-spoofing.

The rest of the Chapter is organised as follows. In Section 7.2, we review the relevant literature. In Section 7.3, we illustrate briefly the methods which are used here in this chapter. Section 7.4 describes the VQ-VAE face anti-spoofing algorithm and the databases we used. In Section 7.5, we present the bias analysis on the SiW database, and in Section 7.6, the bias analysis on the RFW database. We briefly conclude in Section 7.7.

7.2 Related Work

In this section, our primary focus is to delve into related research that directly pertains to in-face anti-spoofing. While Chapter 2 offered a comprehensive overview of the broader field of face anti-spoofing, our attention is now narrowed to studies and investigations that are closely related to bias within this domain. Specifically, we provide a concise review of research related to bias in machine learning and Presentation Attack Detection (PAD). This review aims to shed light on the intricacies of bias within these areas.

7.2.1 Bias in machine learning

The bias in machine learning is a hot topic that is intricately linked to several subjects and areas of study, encompassing disciplines such as ethics, fairness, social sciences, policy, law and data collection [165]. In [166], several high-profile cases of machine learning bias are documented; Google search results appeared to be biased towards women in 2015; Hewlett-Packard's software for web cameras struggled to recognize dark skin tones; and Nikon's camera software was inaccurately identifying Asian people as blinking.

Thus, also given the ethical, legal, and regulatory issues associated with the problem of bias within human populations, there is a considerable amount of research on the subject, especially in face recognition (FR). A recent comprehensive survey can be found in [167], where the significant sources of bias [168, 169] are categorised and discussed, and the negative effect of bias on downstream learning tasks is pointed out. We also note that while the current deep learning-based FR algorithms are

under intense scrutiny for potential bias [170], this is due to their wider deployment in real-life applications rather than any evidence that they are more biased than traditional approaches.

In one of the earliest studies of bias in FR, predating deep learning, [171] reported differences in the performance of humans of Caucasian and East Asian descent between Western and East Asia developed algorithms. In [172], several deep learning-based FR algorithms are analysed, and a small amount of bias is detected in all of them. Then, the authors show how this bias can be exploited to enhance the power of malicious morphing attacks on FR-based security systems.

In [173], the authors compute cluster validation measures on the clusters of the various demographics inside the whole population, aiming at measuring the algorithm's potential for bias. Their result is negative, and they argue for the need for more sophisticated clustering approaches. We note that in our Chapter, an investigation in the latent space of the potential for bias by measuring the discriminative power of SVMs over the various ethnicities returned a similarly negative result. In [174], the aim is to detect bias by analysing the activation ratios at the various layers of the network. Similarly to our work, their target application is the detection of race bias on a binary classification problem, gender classification in their case. Their result is positive in that they report a correlation between the measured activation ratios and bias in the final outcomes of the classifier. However, it is unclear if their method can be used to measure and assess the statistical significance of the expected bias.

In Cavazos et al. [175], similarly to our approach, most of the analysis assumes a one-sided error cost; in their case, the false acceptance rate and the decision thresholds are treated as user-defined variables. However, the analytical tools they used, mostly visual inspection of ROC curves, do not allow for a deep study of the distributions of the similarity scores. At the same time, here, we give a more detailed analysis of the distribution of the responses, which is the equivalent of the similarity scores. In Pereira and Marcel [176], a fairness metric is proposed, which can be optimised over the decision thresholds, but again, there is no in-depth statistical analysis of the scores, as we do here for the responses, and thus they offer more

limited insight.

7.2.2 Bias in Presentation Attack Detection

The literature on bias in presentation attacks is more sparse. Race bias was the key theme in the competition of face anti-spoofing algorithm on the CASIA-SURF CeFA database [177]. Bias was assessed by the algorithm's performance under a cross-ethnicity validation scenario. Standard performance metrics, such as APCER, BPCER and ACER were reported. In [178], the standard CNN models ResNet50 and VGG16, were compared for gender bias against the debiasing-VAE proposed in [179], and several performance metrics were reported. In a recent white paper published by the ID R& D company, a developer of face anti-spoofing software, the paper presented the findings of a comprehensive bias assessment experiment carried out by Bixelab, an independent laboratory accredited by NIST [159].

Similarly to our approach, they focus on bona fide errors, and their aim is for the BPCER error metric to be below a prespecified threshold across all demographics.

Regarding other biometric identification modalities, [180] studied gender bias in iris PAD algorithms. They reported three error metrics, APCER, BPCER, and HTER, finding that female users would be less protected against iris PAD attacks.

7.2.3 Databases

As previously mentioned in Chapter 2, the *Replay-Attack* database [156] comprises 50 subjects from three different ethnicities, with a distribution of 76% Caucasian, 22% Asian, and 2% African. Our second training dataset is sourced from *SiW* [124], which includes 165 subjects representing four ethnicities: 35% of Asian and 35% Caucasian and 23% Indian, and 7% African American. To conduct the bias analysis, we utilized the SiW dataset, where we manually annotated subjects by their ethnicity, and also the pre-annotated RFW database [181].

The first face anti-spoofing database to include explicit ethnic labels was *CASIA-SURF CeFA* [182], which has 1,607 in three ethnicities, captured in three modalities. In this Chapter, for bias analysis, we use the RFW [181], which includes four types

of ethnicities, Caucasian, Asian, Indian, and African. RFW does not specialise in face anti-spoofing and is more widely used in the bias analysis literature.

7.3 Methods

In this section, we will provide a concise overview of the background of the methods used in this chapter. The aim is to briefly explain the underlying concepts and principles behind these methods, providing the necessary context for better comprehension.

7.3.1 Vector Quantized Variational Autoencoder

Vector Quantized Variational Auto-Encoders (VQ-VAEs) [160] is a powerful deep learning architecture to generate and reconstruct images, especially in anomaly context [183, 184]. In the context of face anti-spoofing, VQ-VAE can be applied to face anti-spoofing by combining the strengths of VAEs and vector quantization, enabling the development of a robust and effective face anti-spoofing system by learning discriminative representations of face images.

The VQ-VAE [160] architecture comprises three parts. A learned *codebook* is used to discretise the continuous latent vectors to a set of discrete latent variables; each continuous vector is replaced with its nearest vector in the codebook. An *encoder* that maps the input to a sequence z of discrete codes. A *decoder* transforms the sequence z of discrete vectors back to an image. This quantization process reduces the dimensionality of the representation and, thus, enforces the compactness of the learned representation.

In our context, the VQ-VAE can be used to capture the subtle differences between genuine and fake face images in the context of anomaly detection. By training the VQ-VAE with genuine images only, and optimizing it to minimize the reconstruction error in the class of genuine images, the VQ-VAE can learn to differentiate between genuine and fake face images based on their distinct visual patterns.

Next, we describe in more detail the main components of the VQ-VAE. In a usual variational autoencoder, the encoder captures the posterior distribution $q(z \mid$

x) by mapping the input data x to Gaussian distributions. Then, the decoder models the distribution $p(x \mid z)$ to reconstruct the input data. The VQ-VAE builds upon the common VAE structure but introduces an additional step known as vector quantization (VQ). Instead of using Gaussian distributions, the VQ-VAE applies a categorical distribution in both the posterior and prior distributions. The posterior categorical distribution $q(z \mid x)$ is defined as follows:

$$q(z = k \mid x) = \begin{cases} 1 & \text{for } k = \operatorname{argmin}_j \|z_e(x) - e_j\|_2 \\ 0 & \text{otherwise} \end{cases} \tag{7.1}$$

where, $z_e(x)$ represents the output of the encoder, e_j represents the learned codebook used for the vector quantization of $z_e(x)$, and k is the index of the selected vector in the codebook that best represents $z_e(x)$. The vector-quantized representation $z_q(x)$, which serves as the input for the decoder, is determined through Equations (7.1) and (7.2):

$$z_q(x) = e_k, \quad \text{where} \quad k = \operatorname{argmin}_j \|z_e(x) - e_j\|_2 \tag{7.2}$$

Finally, the output of the encoder $z_e(x)$ is represented by a combination of vectors from the codebook.

The VQ-VAE loss function can be written as the sum of three terms: the first term is the reconstruction loss, which measures the difference between the original data and its reconstruction:

$$L_R = \|x - dec\,(enc(x))\|_2^2 \tag{7.3}$$

where x is the input, $enc(x)$ is the latent representation obtained from the encoder, and $dec(enc(x))$ is the reconstructed data obtained from the decoder. The reconstruction loss is typically calculated using mean squared error (MSE) between the original data and reconstructed data, as in (7.3), but other error metrics could be used too.

The second term is the Codebook loss plays a central role in VQ-VAE. It measures how far the encoded latent vectors are from the nearest codebook entries,

effectively encouraging the model to find the best discrete latent representation for each input. It promotes the learning of meaningful and compact representations.

$$L_{codebook} = \|enc(x) - e\|_2^2 \tag{7.4}$$

The third term is the commitment loss, which measures the difference between the encoder output and its nearest neighbour in the discrete representation of the codebook

$$L_C = \|enc(x) - x_e\|_2^2 \tag{7.5}$$

where e is the nearest code in the discrete representation to the encoder output. The commitment loss encourages the encoder to produce a quantized representation close to the codebook. The overall VQ-VAE loss is combining (7.3), (7.4) and (7.5) after substituting for *enc* and *dec* using (7.1) and (7.2):

$$L = \log(p(x \mid q(x))) + \|\text{sg}\left[z_e(x)\right] - e\|_2^2 + \beta \left\|z_e(x) - \text{sg}[e]\right\|_2^2 \tag{7.6}$$

where sg denotes the stop-gradient operator and β is hyper-parameter which used for adjusting the weight for the commitment loss, that is, the parameter β controls the trade-off between the reconstruction loss and the commitment loss. This VQ-VAE model was used in this chapter to study racial bias in the face of anti-spoofing databases.

7.3.2 Support Vector Machine

Support Vector Machines (SVM) [185], which is a popular machine learning algorithm used for classification and regression tasks, is used in this chapter as an algorithmic test for detecting racial bias in the latent space of the VQ-VAE encoding that is, before a scalar response or a binary outcome have been produced. It is a supervised learning algorithm that analyzes data and builds a model to predict or classify new instances.

The primary concept involves discovering a hyperplane that optimally distinguishes data points from different classes.

That is the computed hyperplane acts as the decision boundary that separates the data into different classes. SVM training processes aim at maximizing the margin, that is, the distance between the hyperplane and the nearest data points from each class. The data points that are closest to the hyperplane and influence its position are the so called support vectors.

7.3.3 Statistic Analysis

Chi-Square Test

The chi-squared test [186] is a statistical test used to determine whether there is a significant association between two categorical or independent variables. It is based on the chi-squared statistic, which measures the difference between the observed and expected frequencies under the null hypothesis that the two categorical variables are independent. The input for the computation of the chi-squared statistic usually has the form of *contingency tables*, which record the frequencies or counts of the different categories of the two variables.

Mann-Whitney U Test

The Mann-Whitney U test [187] is a non-parametric statistical test used to determine if there is a significant difference between the distributions of two independent groups. It is often used as an alternative to the independent samples t-test when the data does not meet the assumptions of normality or when the data are measured on an ordinal or non-continuous scale. The test can provide a p-value, which indicates the probability of observing the data if the null hypothesis is true. If the p-value is below a predetermined significance level (e.g., 0.05), it suggests that there is evidence to reject the null hypothesis and thus conclude that there is a significant difference between the two groups.

The Mann-Whitney U test is widely used in various research fields, including biology, psychology, social sciences, and medicine, where it is often employed to compare the distributions of variables between different groups or conditions. Here, we will use it to analyse the scalar responses of our classification algorithm over the

various races.

Hartigan's dip test

Hartigan's dip test [161], also known as Hartigan's D statistic, is a statistical test used to assess the unimodality of a univariate dataset. It determines whether a given dataset follows a unimodal distribution or exhibits significant departures from unimodality, indicating the presence of bimodality or even multiple modes. Here, we will use it in the analysis of the scalar responses of our classification algorithm, on each race separately.

7.4 Experimental setup

In this section, we briefly describe and validate the two classifiers that will be analysed for bias in the subsequent sections. They have the same VQ-VAE-based architecture [160], and their only difference is in training; in the first, the network is trained on Replay Attack, and in the second on SiW. We chose the VQ-VAE [160] architecture because of some recently reported impressive results on various computer vision problems [188], even though it has not been employed yet in the field of face anti-spoofing.

7.4.1 The VQ-VAE network

In our VQ-VAE, the encoder consists of two convolutional layers with kernel size 4, stride step 2, padding 1, and followed by a ReLU, one convolutional layer with kernel size 3, stride step 1, padding 1 and followed by two residual blocks, which are implemented as ReLU, 3×3 conv, ReLU, 1×1 conv for each block. The decoder is a symmetrical structure of the encoder that uses transposed convolutions. The encoder may output a 16x16 grid of vectors, and each one is quantized, using a codebook of size 512, before being fed to the decoder. The weight factor β was set to 0.25.

The ADAM optimiser was used with a learning rate of 1e-3, and the model was trained for 100 epochs with a batch size of 16. The total loss consisted of three

components: reconstruction loss, codebook loss, and commitment loss. The code was written entirely in Python, on the Pytorch platform. The experiments were run on an Intel Core i7 CPU, with 64 GB of RAM and an Nvidia GTX 1650. Figure 7.1 shows the architecture of the VQ-VAE.

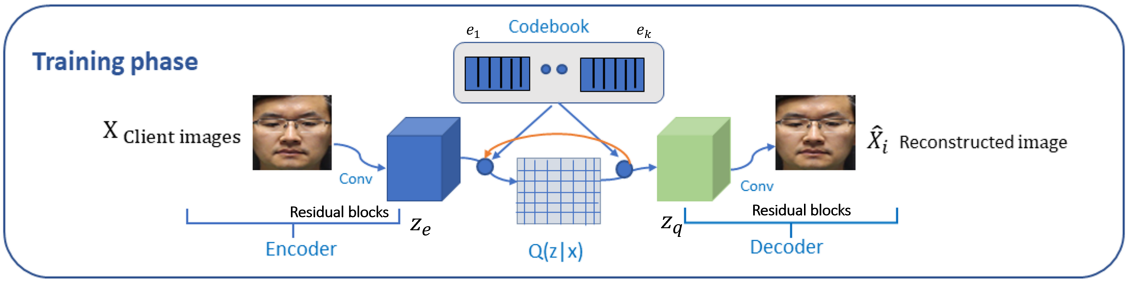

Figure 7.1: The architecture of the proposed VQ-VAE.

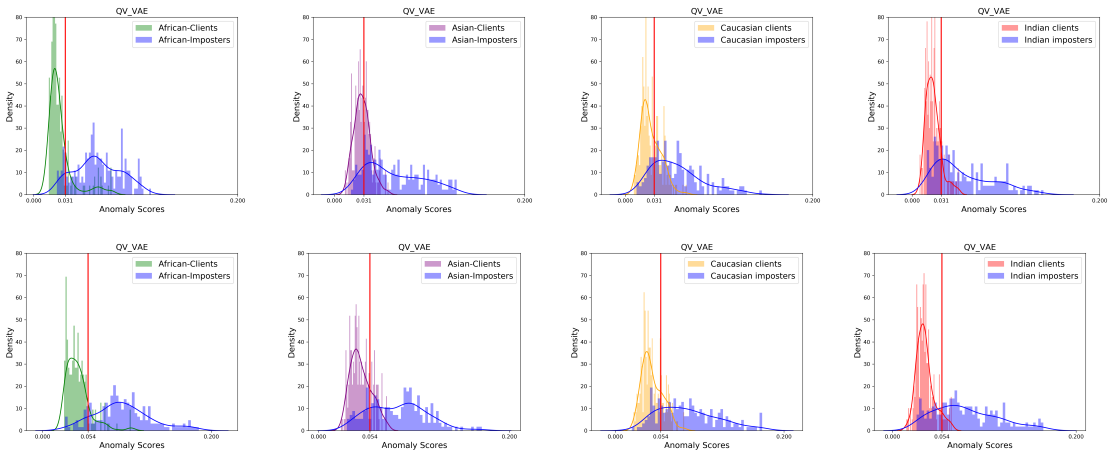

Figure 7.2: **Top:** The responses of the RA trained network on each race of the SiW testset. From left to right: African, Asian, Caucasian, Indian. **Bottom:** The responses of the SiW trained network.

7.4.2 Data preparation

For face detection, we used the Multi-Task Cascade Convolutional Neural Network (MTCNN) [88]. The detected faces were rotationally aligned, making the line between the two eye centres horizontal, and cropped at 64×64 resolution. Table 7.1 summarises the sizes of the datasets. Notice that the training sets consisted of bona fide data, only.

Table 7.1: Sizes of training and test datasets. The test sets are equally split between clients and imposters. The SiW test set, consists of 400 samples from each race.

	Training	**Test**
RA	7465	400
SiW	124,000	1600

7.4.3 Network validation

We validated the classifiers on RA and SiW, first under an intra-database protocol, and then with a cross-database protocol. When testing on SiW, we also report error rates for each race separately.

Intra-database protocol: on an independent validation set we compute the threshold corresponding to the Equal Error Rate (EER). We use this threshold to compute the Half-Total Error Rate (HTER) on the test set. On SiW, we use a single threshold for all races.

Cross-database protocol: here, thresholds estimated on one database do not work on the other. So, we just report the EERs computed directly on the test sets.

Table 7.2 shows the error computed under the testing protocol. The intra-database error is significantly smaller in RA than in SiW (.055 vs .169), indicating an easier classification task, as RA has less diversity in poses and expressions. However, as expected, the error on SiW is smaller when we train on SiW rather than RA (.169 vs .208).

Table 7.2: Error rates computed under the testing protocol.

	RA test	**SiW test**
RA training	.055 [thr = .015 (HTER)]	.208 [thr = .031 (EER)]
SiW training	.180 [thr = .065 (EER)]	.169 [thr = .054 (HTER)]

Table 7.3 shows HTERs for each race of the SiW test set. The thresholds are taken from the corresponding tests of Table 7.2. We note that the SiW trained network is performing better on the SiW test set, not only in total but also on each race separately. For each classifier and each race, Figure 7.2 shows the histograms of the responses and the corresponding thresholds.

Table 7.3: HTERs computed for each race separately on the SiW testset.

	Af	**As**	**Ca**	**In**
HTER RA training (thr = .031)	.135	.247	.260	.192
HTER SiW training (thr = .054)	.115	.202	.210	.150

7.5 Bias analysis on SiW

Here, we use the SiW testset of Section 7.4. The binary outcomes of the classifiers are analysed with the chi-squared test, the scalar responses with the Mann–Whitney U test [187], and finally, the encoding vectors are analysed by using them to train and test an SVM on the task of race classification. The bias analysis process is summarised in Fig. 7.3.

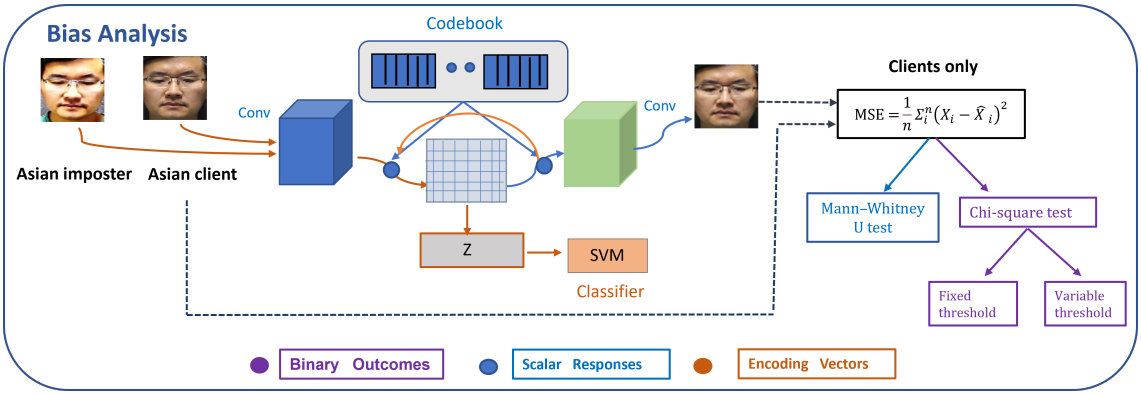

Figure 7.3: The bias analysis process. The binary outcome analysis is shown in purple, the scalar responses analysis in blue, and the latent space analysis in orange.

7.5.1 Statistical analysis of the binary outcomes

First, we analyse the binary outcomes corresponding to the thresholds of Section 7.4.3, that is, 0.031 for the RA-trained classifier and 0.054 for the SiW trained classifier. For each pair of races, we formed the 2×2 contingency tables, and applied the chi-squared test, computing p-values for the hypothesis that the probability of misclassification of a bona fide sample from the race with the most misclassifications is higher.

The results are summarised in Table 7.4. In several cases, the p-values are low, meaning that for any reasonable threshold of statistical significance, the bias hypothesis is accepted. In other cases, p-values above 0.05 mean that bias has not been detected.

Table 7.4: p-values of the chi-squared tests for the thresholds in Section 7.4.3

	Af-As	**Af-Ca**	**Af-In**	**As-Ca**	**As-In**	**Ca-In**
RA training	.0000	.0001	.1465	.2532	.0000	.0000
SiW training	.1158	.0104	.0147	.0000	.0000	1.0

Next, we treat the thresholds, that is, the operating points on the ROC curve, as a variable. In Figs. 7.4, 7.5, we plot the p-value as a function of the response, for the two classifiers and the six pairs of races. We note that, over the range of thresholds,

there are several disconnected intervals corresponding to high bias, which means that threshold optimisation for low bias should not assume unique solutions, as it is often implicit in the literature.

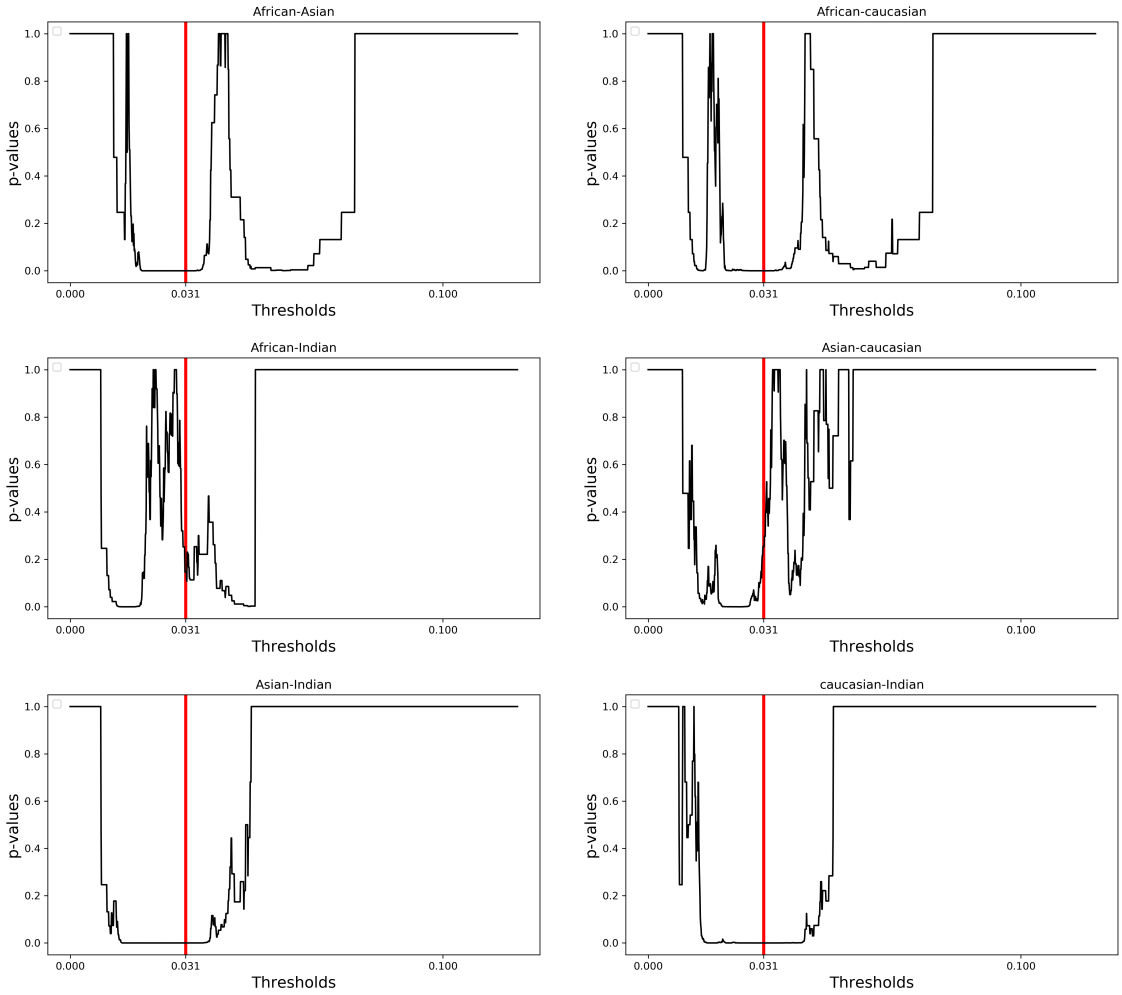

Figure 7.4: For each pair of races, graphs of the p-value as a function of the threshold. The classifier was trained on RA.

7.5.2 Statistical analysis of the scalar responses

For an insight in the behaviour of the graphs of the p-values we analyse the scalar responses of the algorithm. Figs. 7.6, 7.7 show a comparison of the histograms of the responses for each pair of races. We note the complex behaviour of the density functions, which induce a complex bias behaviour. In particular, bias at certain thresholds can be caused by differences in the means of the responses; differences in their variances; response bimodality; or outliers.

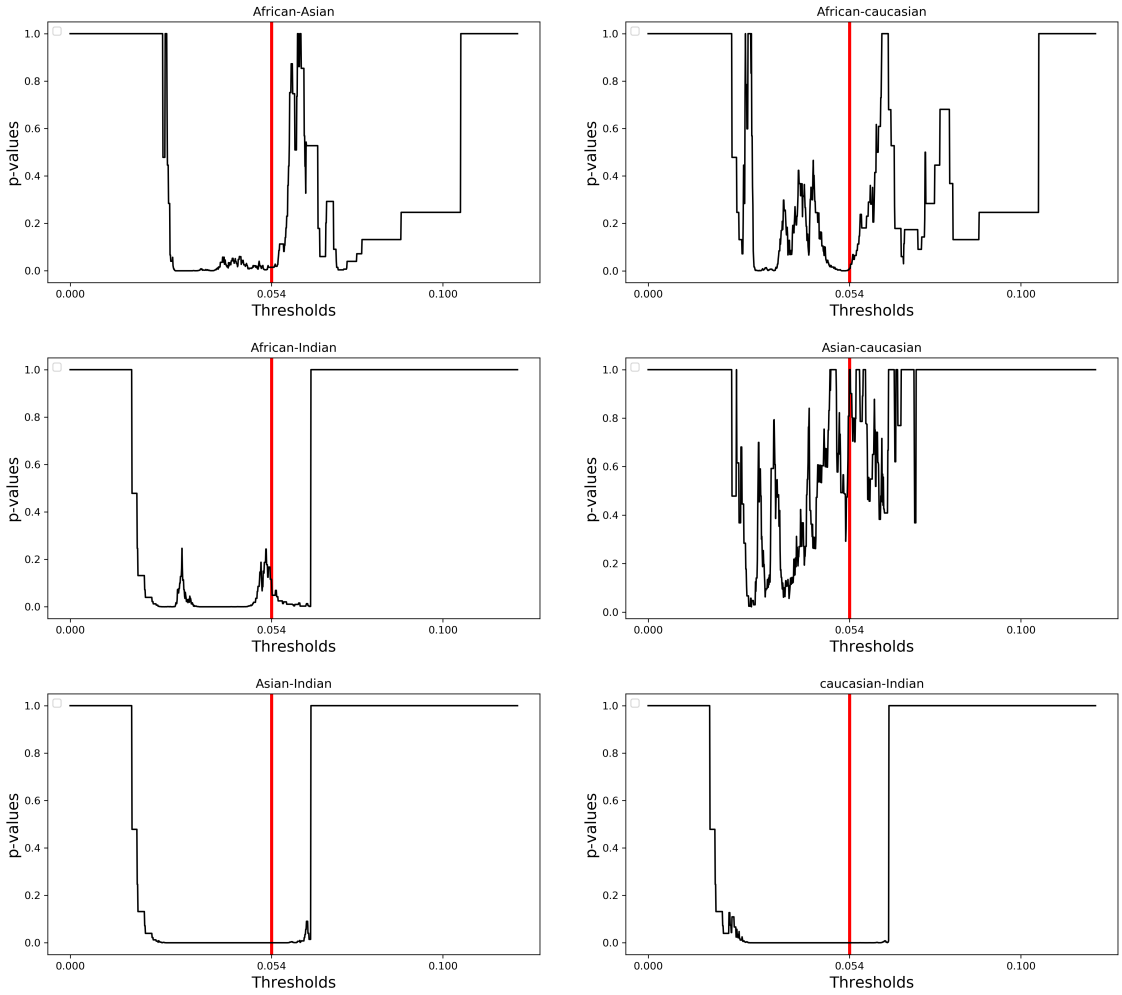

Figure 7.5: For each pair of races, graphs of the p-value as a function of the threshold. The classifier was trained on SiW.

Regarding the means of the responses, since the Shapiro-Wilk test rejected the hypothesis of normal distributions, we opted for the Mann–Whitney U test. Table 7.5 shows the computed p-values on each pair of races, for the hypothesis that the values of randomly selected responses from the two populations are different.

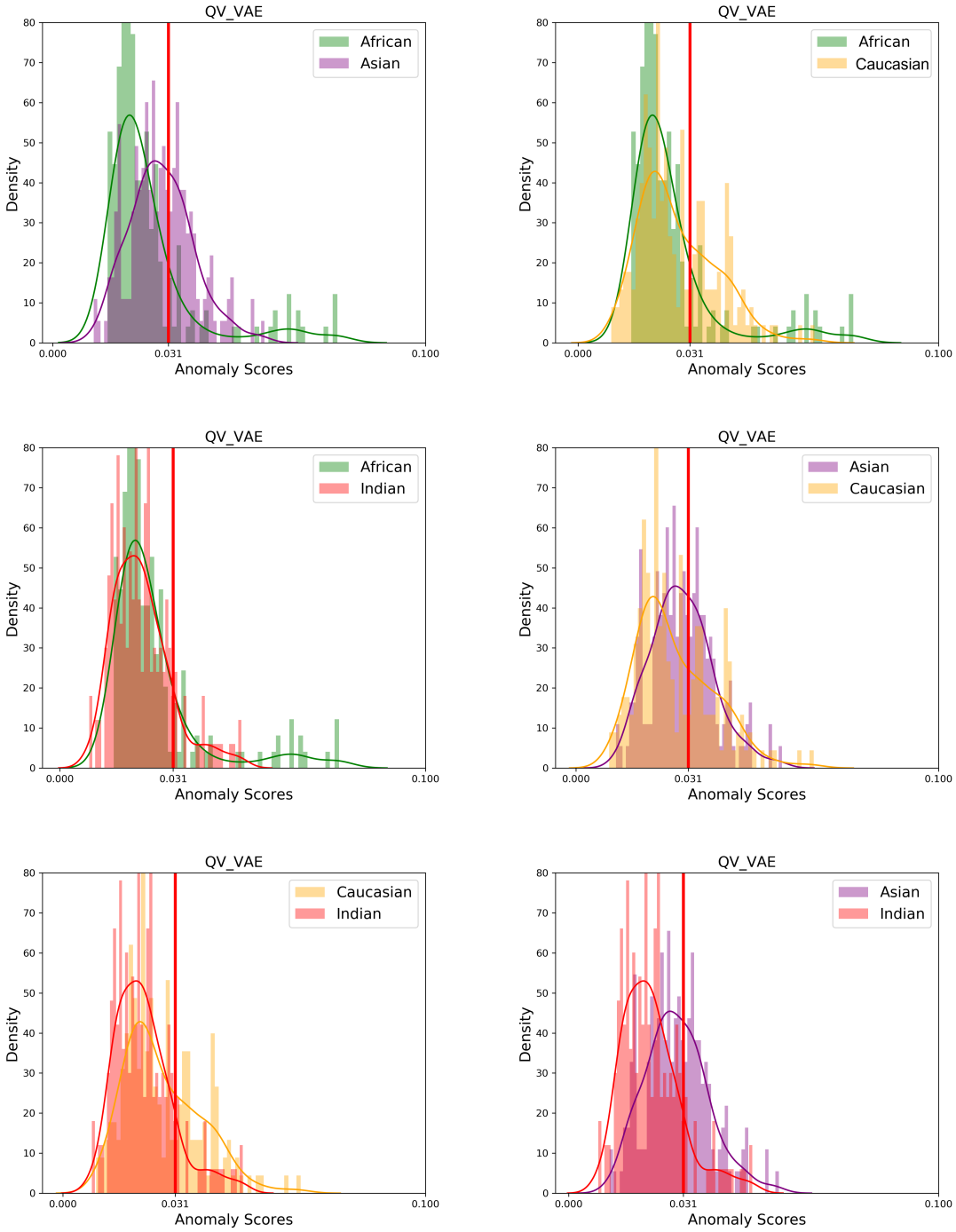

Figure 7.6: For each pair of races in SiW, the histogram of the responses on bona-fide images. The classifier was trained on RA.

Table 7.5: p-values of the Mann–Whitney U test on each pair of races.

	Af-As	**Af-Ca**	**Af-In**	**As-Ca**	**As-In**	**Ca-In**

Continued on next page

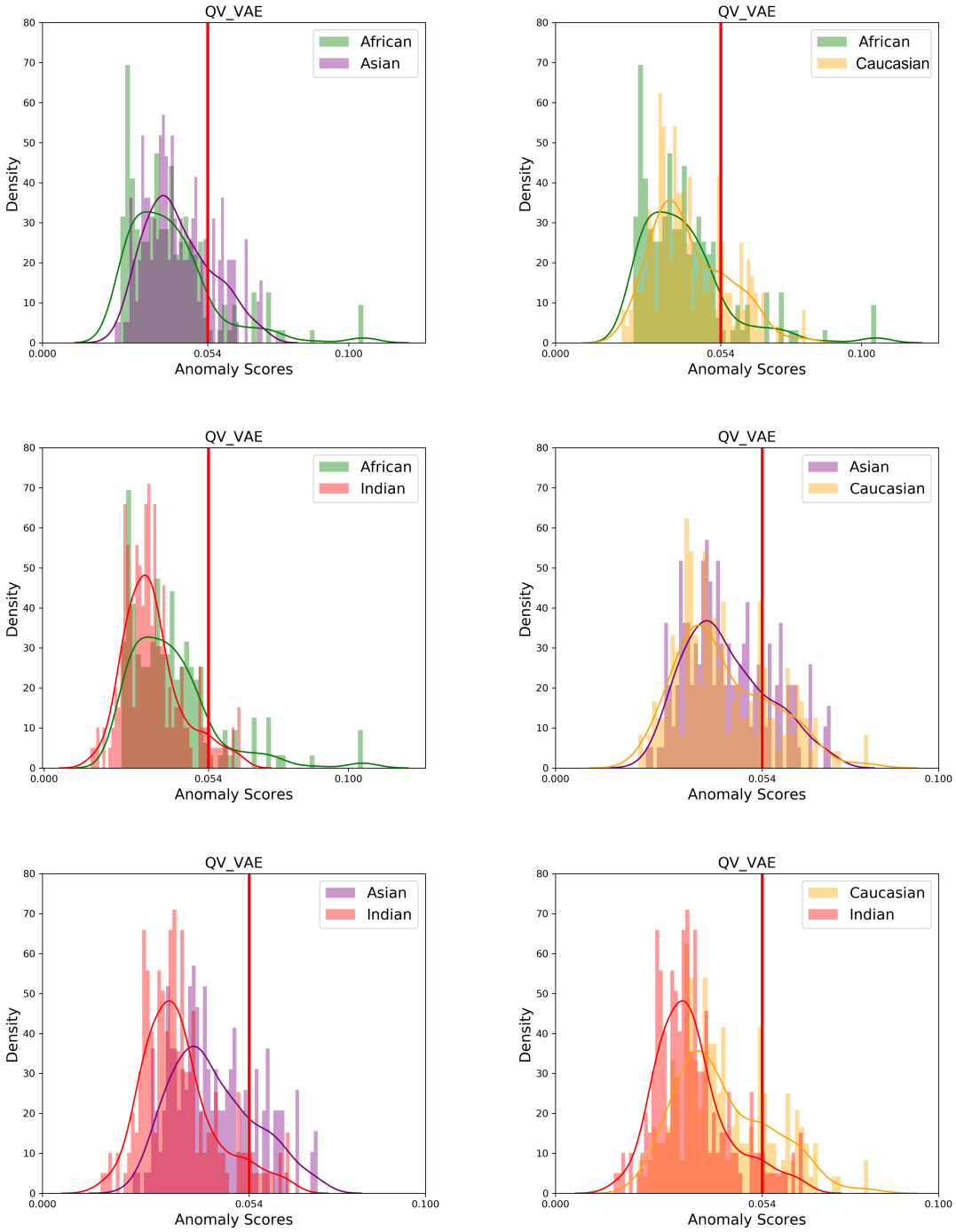

Figure 7.7: For each pair of races in SiW, the histogram of the responses of bona-fide images. The classifier was trained on SiW.

Table 7.5: p-values of the Mann–Whitney U test on each pair of races. (Continued)

RA	.0000	.0015	.0085	.0052	.0000	.0000
training						

Continued on next page

Table 7.5: p-values of the Mann–Whitney U test on each pair of races. (Continued)

SiW training	.0001	.0078	.0000	.1560	.0000	.0000

Regarding bimodality, we used Hartigan's Dip Test [161], with 50 bins, to test each race for bimodality. We note that for the 200 samples we have from each race, a statistical significance of 95% corresponds to a critical value of 0.037. Table 7.6 shows the computed dip values, together with the means and standard deviations for each population.

We note that the biggest difference in mean response is between Asian and Indian from the SiW trained classifier, while the corresponding standard deviations are similar. Thus, the bias we observe in the corresponding diagram in Fig 7.5 is due to the classifier's higher mean response on Asians compared to Indians. The substantial difference in mean response between Asian and Indian populations from the SiW trained classifier is likely due to several factors. First, the diverse facial features among these groups, influenced by genetics, ethnicity, and regional characteristics, contribute to variations in their representation within the classifier. Additionally, potential classifier bias, and the quality of the training data can all affect performance and lead to these mean response differences.

On the other hand, the smallest difference in mean response is between Asians and Caucasians, again under the SiW trained classifier. Thus, the bias we observe in the corresponding diagram in Fig 7.5, which for very small threshold values is statistically significant, is due to different variances. We note here, that while at such low threshold values, a standalone classifier wouldn't be very useful, it is still an interesting case when constructing ensembles of weak classifiers.

Finally, we notice that all Hartigan's Dip Test values are below the significance threshold, and thus, all populations should be considered unimodal. In particular, that means that some very high responses on African people, especially from the RA trained classifier, should be treated as outliers. We note that, nevertheless, against all the other three races, these outliers create a second, or third region of high bias, and in those regions samples from the African population are treated less favourably.

Table 7.6: SiW testset: means, standard deviations, and Hartigan's dip values for the responses of the RA and SiW trained classifiers.

	Af	**As**	**Ca**	**In**
RA training μ	.0257	.0291	.0274	.0223
SiW training μ	.0418	.0446	.0438	.0355
RA training *s.d.*	.0126	.0084	.0104	.0079
SiW training *s.d.*	.0142	.0109	.0122	.0096
RA training dip	.0221	.0360	.0335	.0349
SiW training dip	.0299	.0233	.0366	.0324

7.5.3 Discrete latent space

The proposed VQ-VAE encodes each image with a 32768 number long sequence of integers in [0..1023], corresponding to indices of the vectors in the codebook. Using these latent space representations of the images as inputs, for each pair of races, we train an SVM [185] with an RBF kernel of Gaussian type. Table 7.7 shows AUC values for these SVM classifiers.

Table 7.7: AUC values for an SVM trained on the VQ-VAE's latent space encodings of the images. SiW testset.

	Af-As	**Af-Ca**	**Af-In**	**As-Ca**	**As-In**	**Ca-In**
RA training	.88	.87	.74	.71	.78	.83
SiW training	.74	.85	.76	.72	.76	.83

We note that in some cases the SVMs perform quite well, meaning that the VQ-VAE encodings of the face images contain enough information to discriminate between races. This suggests that the SVM can be employed for classification tasks, including the differentiation of various races based on the latent space, which captures the fundamental features extracted by the VQ-VAE. Despite the fact that VQ-VAEs themselves were not trained with race labels. That means that there is a potential for biased behaviour, however, on the other hand, the statistical significance of the result and its practical implications are not clear, and a deeper investigation is required.

7.6 Bias analysis on RFW

In this section, we apply the same bias analysis on a testset from the RFW database, consisting of 200 images from each race. We note that this time the race labels come as part of the database, rather than being annotated by us. Another difference from SiW is that the RFW database is not a specialised face anti-spoofing database. Thus, as we do not have imposter images, we do not have empirically established thresholds, as for example the ones corresponding to EER values. For the part of the analysis requiring specific thresholds, we use thresholds corresponding to some predetermined TPR values.

Statistical analysis of the binary outcomes

Table 7.8 shows the thresholds corresponding to TPR values of 1%, 2%, 5%, 10%, 20%.

Table 7.8: Threshold values corresponding to some predetermined TPR values.

	1%	**2%**	**5%**	**10%**	**20%**
RA training	.0460	.0440	.0389	.0336	.0279
SiW training	.1206	.1079	.0971	.0838	.0712

Table 7.9 shows the results of the chi-squared tests on the binary outcomes. We note that bias can be detected for some of the thresholds, for some pairs of races. In Figs. 7.8, 7.9, for each classifier, and for each pair of races, we show the histograms of responses on bona fide images, and the p-values of the chi-squared test as a function of the threshold. We observe behaviours similar to those from the tests on the SiW database.

Table 7.9: p-values of the chi-squared tests for the thresholds shown in Table 7.8.

	Af-As	**Af-Ca**	**Af-In**	**As-Ca**	**As-In**	**Ca-In**
RA training 1%	1.0	.0718	.2171	.1316	.3680	1.0
RA training 2%	1.0	.1271	.2839	.0741	.1774	1.0
RA training 5%	.8494	.0299	.1082	.0116	.0483	.7487
RA training 10%	.0356	.0069	.3938	.0000	.0021	.0857
RA training 20%	.1333	.0025	1.0	.0000	.1047	.0037
SiW training 1%	1.0	1.0	1.0	1.0	.3680	.6153
SiW training 2%	.2839	.4456	1.0	1.0	.5001	.7209
SiW training 5%	.2924	.0249	.4143	.31053	1.0	.2152
SiW training 10%	.1158	.0302	.3567	.6501	.6246	.2725
SiW training 20%	.0236	.0062	.01241	.7194	0.9039	.9054

7.6.1 Statistical analysis of the scalar responses

Table 7.10 shows the p-values of the Mann–Whitney U test for each pair of races, and Table 7.11 shows the means, standard deviations and dip values for each population. We note that Hartigan's Dip Test detects a bimodality in the responses of the SiW trained algorithm on Indian people, having a dip value of 0.055, above the

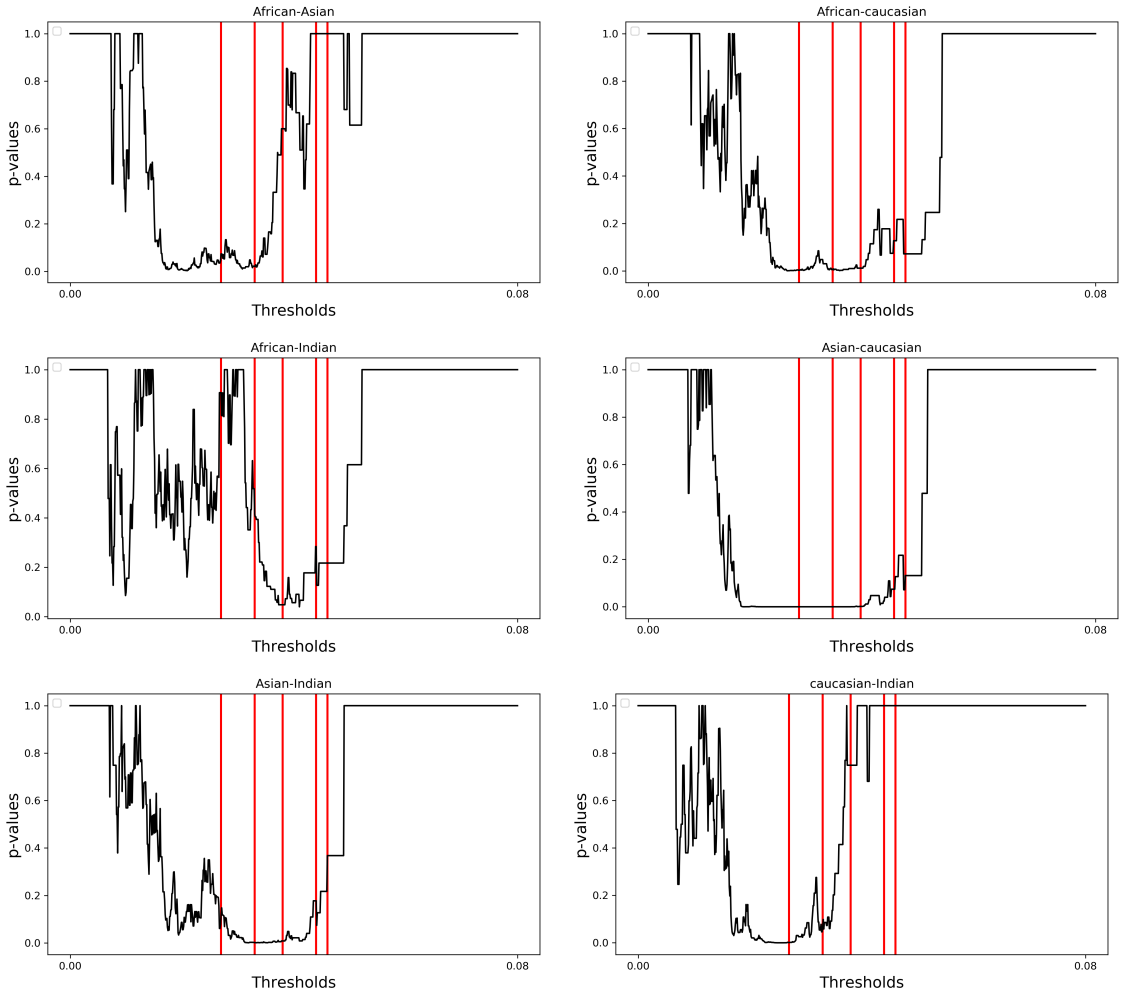

Figure 7.8: For each pair of races, graphs of the p-value as a function of the threshold. The classifier was trained on RA.

significance threshold of 0.037. This can also be verified by visual inspection of the corresponding histograms of responses, which, for each pair of races, are shown in Figs. 7.10, 7.11. We also note that this bimodality can be detected in the behaviour of corresponding graphs of the p-values of the chi-squared test. Indeed, in the three graphs in Fig. 7.9 corresponding to Indian people, we can detect two distinct regions of higher bias, even though the second one does not reach the level of statistical significance. We observe that for the first two rows in both classifiers at the specified threshold values, which correspond to TPR (True Positive Rate) values of 1% and 2%, there is no bias.

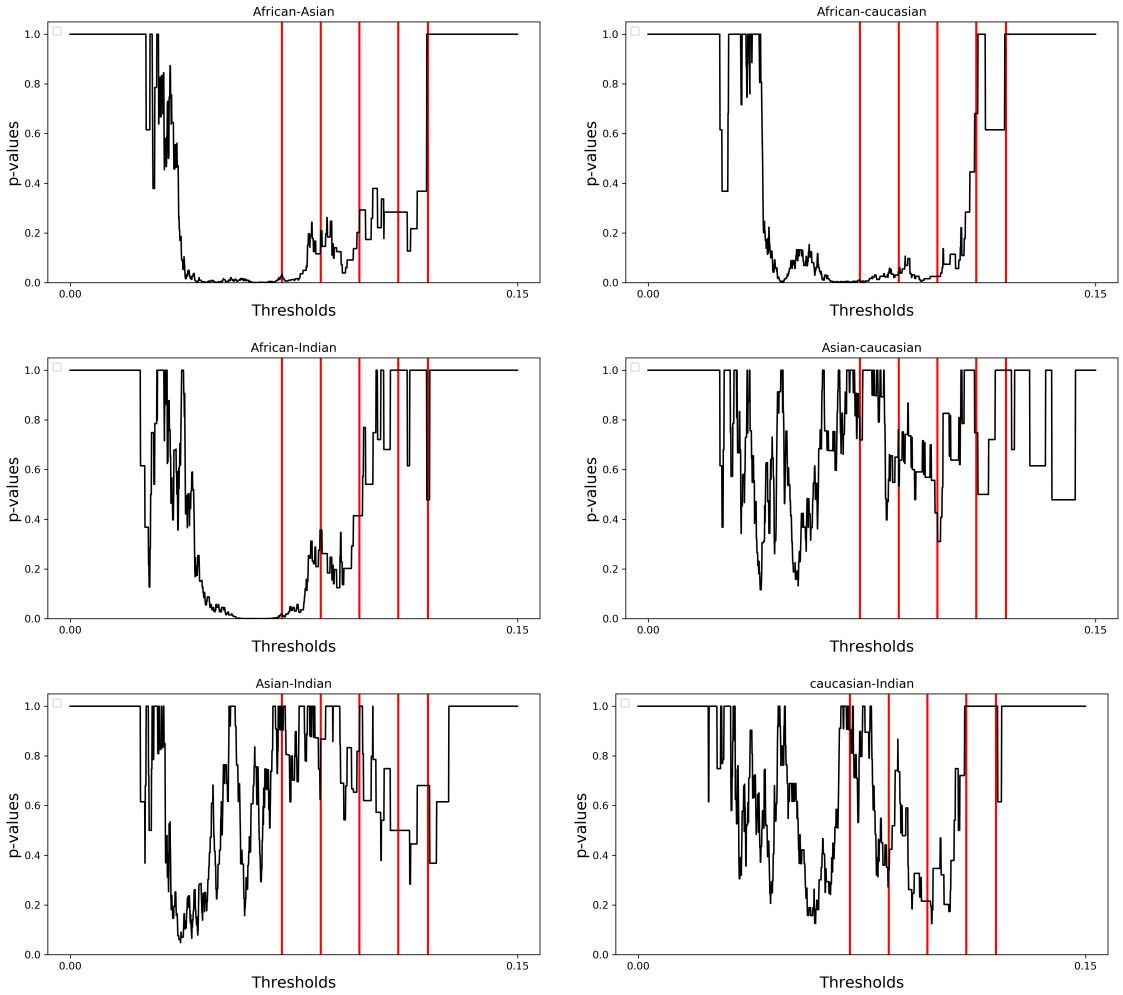

Figure 7.9: For each pair of races, graphs of the p-value as a function of the threshold. The classifier was trained on SiW.

Table 7.10: p-values of the Mann–Whitney U test on each pair of races.

	Af-As	**Af-Ca**	**Af-In**	**As-Ca**	**As-In**	**Ca-In**
RA training	.0067	.0253	.3216	.0000	.0137	.0044
SiW training	.0004	.0058	.0062	.2509	.2743	.4805

Table 7.11: RFW testset: means, standard deviations, and Hartigan's dip values for the responses of the RA and SiW trained classifiers.

	Af	**As**	**Ca**	**In**
RA training μ	.0217	.0239	.0194	.0214
SiW training μ	.0509	.0579	.0569	.0579
RA training $s.d.$	.0094	.0101	.0068	.0082
SiW training $s.d.$	.0175	.0220	.0223	.0220
RA training dip	.0175	.0178	.0299	.0216
SIW training dip	.0114	.0225	.0149	.0550

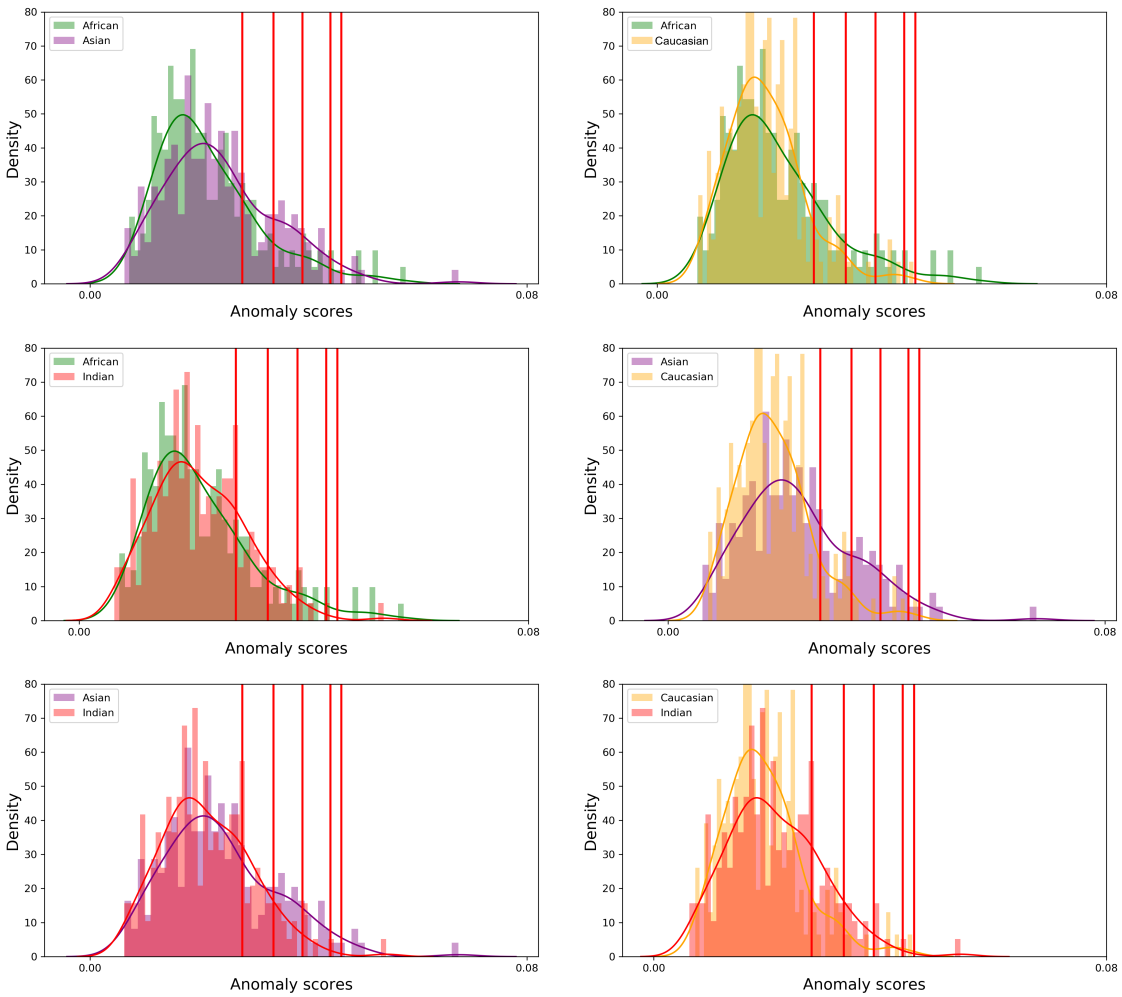

Figure 7.10: For each pair of races in RFW, the histogram of the responses of bona-fide images. The classifier was trained on RA.

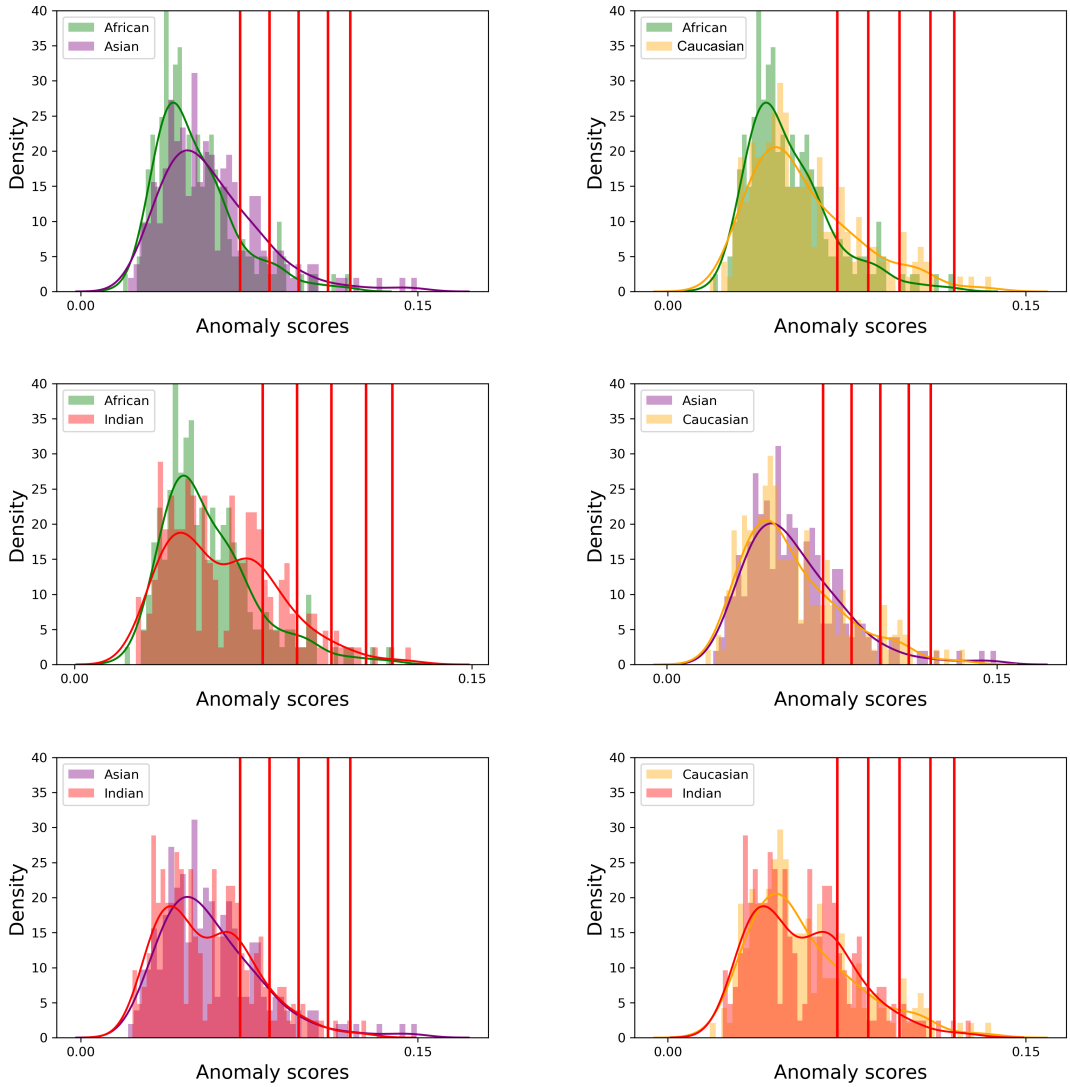

Figure 7.11: For each pair of races in RFW, the histogram of the responses of bona-fide images. The classifier was trained on SiW.

7.6.2 Discrete latent space

Table 7.12 shows the AUC values of SVM classifiers trained and tested, on each pair of races, with the Quantised Vectors. We note an even better performance than in the case of the SiW test set, implying an even higher potential for bias. We can observe that the SVM can effectively distinguish African features from others in both classifiers.

Table 7.12: AUC values for an SVM trained on the VQ-VAE's latent space encodings of the images. RFW testset.

	Af-As	**Af-Ca**	**Af-In**	**As-Ca**	**As-In**	**Ca-In**
RA training	.95	.97	.94	.89	.89	.86
SiW training	.96	.98	.94	.89	.86	.85

7.7 Conclusion

We conducted a systematic, largely empirical, study of race bias in face anti-spoofing, based on a series of statistical and algorithmic tests. Our premises were that we are interested in the bona fide error; that binary outcomes, scalar responses, and latent space should all be analysed for bias; and that the threshold determining the classifier's operating point should not be assumed fixed by a black-box procedure.

Our main finding is that the behaviour of race bias depends on several characteristics of the distributions of the corresponding classifier's responses: differences in the means; differences in the variance; bimodal responses; outliers. The main implication of this finding is that the behaviour of bias can be quite complex, and, for example, methods optimising for low race bias should not assume unique solutions to the problem. More generally, in our context, bias should be treated as something more complex than the difference in the means of two populations, a misconception that might be reinforced by the fact that in statistics, colloquially, the term bias is often used to describe the component of the error attributed to the difference in the means.

CHAPTER 8

Conclusion

8.1 Introduction

Biometric security is the authentication of an individual's identity via an immediate and automatic verification process that uses the user's attributes to grant access to the desired system. Biometric security is currently one of the most reliable identity verification methods, whereby the biometric characteristics employed can be physiological or behavioural. The most effective physiological biometrics types are the face, iris and fingerprint, while behavioural biometrics include activities such as walking, gesturing, and voice. Facial recognition is considered the most prevalent identification technology at a commercial level, and it is a very popular option for security systems. However, it has been well-documented that biometrics are susceptible to malicious attacks. These types of spoofing attempts can be categorised as either direct or indirect. Direct attacks target the system's sensors directly without requiring any prior knowledge about the system's underlying architecture. Direct attacks can be launched by presenting the system's sensor with printed photos or replaying videos of the biometric modalities. Indirect attacks, on the other hand, need knowledge of the computational approaches utilised by the system. Our research

focuses on direct attacks, and in particular, on using anomaly detection methods for face anti-spoofing.

The contributions of this thesis are summarised in Section 8.2. The limitations of the proposed methods are outlined in Section 8.3, and the prospects for future work are discussed in Section 8.5.

8.2 Research contributions

The contributions of the research, as documented in this thesis, are summarised as follows:

1. Creating enhanced presentation attacks.

   Testing the performance of a well-known FAS algorithm against presentation attacks with images that have undergone colour manipulation. The main contribution of Chapter 4 is an evaluation of the sensitivity of ResNet50, which has been evaluated against familiar face anti-spoofing attacks in [59], against processed imposter image attacks from Replay-Attack DB. In a real-world scenario, it is expected that attackers will likely attempt to process the images intended for spoofing to enhance the effectiveness of their attacks.

   The experiments were done in two different ways. Firstly, we propose a colour manipulation adversarial attack on a FAS. To the best of our knowledge, it is the first study of adversarial attacks on deep neural networks in the context of FAS. Secondly, we experimented, the result of which indicates that the proposed adversarial attack can be converted into a direct presentation attack. Our results showed that the accuracy rate of the FAS is affected by the use of processed imposter images.

2. Construction of training sets for anomaly detection.

   Evaluating anomaly detection methods after augmenting the training data set with wild images and images from a non-specialised database and a face anti-spoofing database. In Chapter 5 of the thesis, wild images were aggregated with images from non-specialised DBs to evaluate the performance of an

anti-spoofing method based on anomaly detection, aiming at improving generalisation. We developed an anomaly detection method for face anti-spoofing based on a convolutional autoencoder. We tested it on the previously unseen NUAA database, showing performance increases when we added in-the-wild face images and face images from non-specialised databases into the training set. It was subsequently found that the performance under a cross-database testing protocol increased when the model was trained with the augmented training set.

3. Proposing a novel model for FAS based on anomaly detection.

   In Chapter 6, we proposed a novel Anomaly detection (AnoFormer) that combines the strength of transformers and ResNet as backbones to extract features. Then, a comparison of various one-class classifiers: One class SVM, Isolation Forest, Anoformer + FId, and Anoformer+MSE. However, The results show that a decoder with MSE as a loss function outperforms the other configurations. Also, we conducted an ablation study showing that using a combination of ViT and ResNet in the backbone outperforms using single networks. Thanks to the transformer's attention maps which could distinguish between the imposter's and clients' features well. At the end of this chapter, we compared the Anoformer results with our baseline (transformer with MLP), and we found that the anomaly detection model outperforms the two-class model in the RA/SiW cross-database testing, as well as in the SiW/RA cross-database testing in two of the three attack modalities.

4. Racial bias in anomaly detection face anti-spoofing.

   In Chapter 7, racial bias in face anti-spoofing was analysed on an anomaly detection model (VQ-VAE). We conducted a systematic, largely empirical, study of race bias in face anti-spoofing based on a series of statistical and algorithmic tests. Our premises are: that we are interested in bona fide error, whereby binary outcomes, scalar responses, and latent space should all be analysed for bias, and the threshold determining the classifier's operating point should not be assumed to be fixed by a black box procedure. In our study, we

found that race bias is significantly affected by the distribution of the responses of the corresponding classifier: mean difference; variance difference; bimodal response; outliers.

8.3 Limitations

The limitations of the research documented in this thesis are summarized as follows:

1. First contribution's limitation:

   Due to the laboriousness of the task of creating a database with enhanced presentation attacks, the experiment corresponding to the second contribution of Chapter 4 was limited in scope. The creation of a sizeable database with enhanced presentation attacks is left as future work.

2. Limitation from first to fourth contributions:

   Evaluation: The set of performance metrics used to evaluate the effectiveness of the models and methods proposed in this work against spoofing attacks may not have been comprehensive enough. Our choices had often been convenience choices for the sake of direct comparability with existing works. However, the combinations of evaluation metrics and databases they are applied on vary considerably among various works. Thus, in many cases, our choice of comparator methods was limited.

8.4 Privacy Implications of Cloud-Based Face Anti-Spoofing

The rapid adoption of cloud-based computing services has revolutionized various sectors, including face anti-spoofing. However, with the increasing prevalence of these services, it is essential to address the accompanying privacy concerns, particularly when handling sensitive biometric data. This focuses on the privacy challenges associated with running face anti-spoofing frameworks in the cloud, using face anti-spoofing databases as a case study. Facial data used in anti-spoofing is

highly sensitive and often contains personally identifiable information, making data security a top priority. Key privacy measures include data encryption, access controls, data classification, privacy impact assessments, data residency compliance, data minimization, privacy policies, audit systems, data portability, and a privacy-centered design approach. These measures are crucial for maintaining data privacy and complying with privacy regulations

8.5 Future Work

Our recommendations for future work are summarized as follows:

1. Evaluating the effectiveness of our proposed approach for different types of spoofing attacks. Here, we have focused on a particular type of spoofing attack, i.e., presentation attacks using printed or digital images. However, evaluating our approach to other kinds of attacks, such as 3D masks and video replay attacks, would be valuable.

2. We plan to create a database with imposter images corresponding to processed clients' images for presentation attacks.

3. We would like to test the Anoformer on more databases and even use test bona fide images from outside the specialised PAD databases. Our aim would be to demonstrate the generalisation power of anomaly detection further. Moreover, we would like to further develop this model by incorporating new deep-learning techniques.

4. Harvest a large selection of online in-the-wild images for a new database with client class only to be used for training anomaly models. There should be a variety of environments and backgrounds, and different types of cameras will capture the images.

5. Finally, we also plan to develop anomaly detection classifiers based on adversarial models, such as BiGANs [189], AnoGANs [190], which have been employed to tackle the general PAD problem [191] in a two-class training setting.

Bibliography

[1] J. Galbally, J. Fierrez, and J. Ortega-García, "Vulnerabilities in biometric systems: Attacks and recent advances in liveness detection," *Database*, vol. 1, no. 3, pp. 1–8, 2007.

[2] J. Määttä, A. Hadid, and M. Pietikäinen, "Face spoofing detection from single images using micro-texture analysis," in *2011 international joint conference on Biometrics (IJCB)*, pp. 1–7, IEEE, 2011.

[3] Z. Boulkenafet, J. Komulainen, and A. Hadid, "Face anti-spoofing based on color texture analysis," in *2015 IEEE international conference on image processing (ICIP)*, pp. 2636–2640, IEEE, 2015.

[4] Y. Atoum, Y. Liu, A. Jourabloo, and X. Liu, "Face anti-spoofing using patch and depth-based cnns," in *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pp. 319–328, IEEE, 2017.

[5] X. Tan, Y. Li, J. Liu, and L. Jiang, "Face liveness detection from a single image with sparse low rank bilinear discriminative model," in *European Conference on Computer Vision*, pp. 504–517, Springer, 2010.

[6] I. Chingovska, A. Anjos, and S. Marcel, "On the effectiveness of local binary patterns in face anti-spoofing," in *2012 BIOSIG-proceedings of the international conference of biometrics special interest group (BIOSIG)*, pp. 1–7, IEEE, 2012.

[7] Y. Liu, A. Jourabloo, and X. Liu, "Learning deep models for face anti-spoofing: Binary or auxiliary supervision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 389–398, 2018.

[8] Y. Baweja, P. Oza, P. Perera, and V. M. Patel, "Anomaly detection-based unknown face presentation attack detection," in *2020 IEEE International Joint Conference on Biometrics (IJCB)*, pp. 1–9, IEEE, 2020.

[9] C. Chen, Y. Jing, X. Lu, W. Yuan, and L. Ma, "Spoof face detection via semi-supervised adversarial training," in *2022 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–6, IEEE, 2022.

[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2020.

[11] R. Bresan, A. da Silva Pinto, A. Rocha, C. Beluzo, and T. Carvalho, "Facespoof buster: a presentation attack detector based on intrinsic image properties and deep learning," *arXiv*, vol. 1902.02845, 2019.

[12] S. Prabhakar, S. Pankanti, and A. Jain, "Biometric recognition: security and privacy concerns," *IEEE Security & Privacy*, vol. 1, no. 2, pp. 33–42, 2003.

[13] M. Faundez-Zanuy, "Biometric security technology," *IEEE Aerospace and Electronic Systems Magazine*, vol. 21, no. 6, pp. 15–26, 2006.

[14] M. Abreu and M. Fairhurst, "An empirical comparison of individual machine learning techniques in signature and fingerprint classification," in *Biometrics and Identity Management: First European Workshop, BIOID 2008, Roskilde, Denmark, May 7-9, 2008. Revised Selected Papers 1*, pp. 130–139, Springer, 2008.

[15] L. Ballard, D. Lopresti, and F. Monrose, "Forgery quality and its implications for behavioral biometric security," *Trans. Sys. Man Cyber. Part B*, vol. 37, no. 5, 2007.

[16] A. Jain, A. Ross, and S. Pankanti, "Biometrics: a tool for information security," *IEEE Transactions on Information Forensics and Security*, vol. 1, no. 2, pp. 125–143, 2006.

[17] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM computing surveys (CSUR)*, vol. 35, no. 4, pp. 399–458, 2003.

[18] J. Galbally, S. Marcel, and J. Fierrez, "Biometric antispoofing methods: A survey in face recognition," *IEEE Access*, vol. 2, pp. 1530–1552, 2014.

[19] T. Matsumoto, H. Matsumoto, K. Yamada, and S. Hoshino, "Impact of artificial "gummy" fingers on fingerprint systems," in *Optical security and counterfeit deterrence techniques IV*, vol. 4677, pp. 275–289, SPIE, 2002.

[20] D. Nguyen and Q. Bui, "Your face is not your password," *BlackHat DC*, vol. 12, 2009.

[21] L. Omar and I. Ivrissimtzis, "Evaluating the resilience of face recognition systems against malicious attacks.," BMVA Press, 2015.

[22] "Luxand facesdk." `http://www.luxand.com`, Retrieved on October 16th 2023 from.

[23] "About keylemon." `http://www.keylemon.com`, Retrieved on October 16th 2023 from.

[24] "Facelock for apps." `http://www.facelock.mobi/`, Retrieved on October 16th 2023 from.

[25] "Android faceunlock." `http://www.androidcentral.com`, Retrieved on October 16th 2023.

[26] Z. Yu, X. Li, J. Shi, Z. Xia, and G. Zhao, "Revisiting pixel-wise supervision for face anti-spoofing," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 3, pp. 285–295, 2021.

[27] S. R. Arashloo, J. Kittler, and W. Christmas, "An anomaly detection approach to face spoofing detection: A new formulation and evaluation protocol," *IEEE Access*, vol. 5, pp. 13868–13882, 2017.

[28] Z. Yu, Y. Qin, X. Li, C. Zhao, Z. Lei, and G. Zhao, "Deep learning for face anti-spoofing: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–22, 2022.

[29] Z. Ming, M. Visani, M. M. Luqman, and J.-C. Burie, "A survey on anti-spoofing methods for facial recognition with rgb cameras of generic consumer devices," *Journal of Imaging*, vol. 6, no. 12, p. 139, 2020.

[30] R. Cai, H. Li, S. Wang, C. Chen, and A. C. Kot, "Drl-fas: A novel framework based on deep reinforcement learning for face anti-spoofing," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 937–951, 2020.

[31] A. Liu, Z. Tan, J. Wan, Y. Liang, Z. Lei, G. Guo, and S. Z. Li, "Face anti-spoofing via adversarial cross-modality translation," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 2759–2772, 2021.

[32] L. Omar and I. Ivrissimtzis, "Resilience of luminance based liveness tests under attacks with processed imposter images," in *24th WSCG*, pp. 79–82, 2016.

[33] H. Feng, Z. Hong, H. Yue, Y. Chen, K. Wang, J. Han, J. Liu, and E. Ding, "Learning generalized spoof cues for face anti-spoofing," *arXiv:2005.03922*, 2020.

[34] Z. Li, H. Li, K.-Y. Lam, and A. C. Kot, "Unseen face presentation attack detection with hypersphere loss," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2852–2856, IEEE, 2020.

[35] S. Fatemifar, S. R. Arashloo, M. Awais, and J. Kittler, "Client-specific anomaly detection for face presentation attack detection," *Pattern Recognition*, vol. 112, p. 107696, 2021.

[36] A. Jain, A. Ross, and S. Prabhakar, "An introduction to biometric recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 1, pp. 4–20, 2004.

[37] A. K. Jain, A. Ross, and S. Prabhakar, "An introduction to biometric recognition," *IEEE Transactions on circuits and systems for video technology*, vol. 14, no. 1, pp. 4–20, 2004.

[38] Y. Kortli, M. Jridi, A. Al Falou, and M. Atri, "Face recognition systems: A survey," *Sensors*, vol. 20, no. 2, p. 342, 2020.

[39] M. Turk and A. Pentland, "Face recognition using eigenfaces," in *Proceedings. 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 586–591, 1991.

[40] H. J. Seo and P. Milanfar, "Face verification using the lark representation," *IEEE Transactions on Information Forensics and Security*, vol. 6, no. 4, pp. 1275–1286, 2011.

[41] S. R. Arashloo and J. Kittler, "Class-specific kernel fusion of multiple descriptors for face verification using multiscale binarised statistical image features," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 12, pp. 2100–2109, 2014.

[42] I. L. Kambi Beli and C. Guo, "Enhancing face identification using local binary patterns and k-nearest neighbors," *Journal of Imaging*, vol. 3, no. 3, p. 37, 2017.

[43] H. Cho, R. Roberts, B. Jung, O. Choi, and S. Moon, "An efficient hybrid face recognition algorithm using pca and gabor wavelets," *International Journal of Advanced Robotic Systems*, vol. 11, no. 4, p. 59, 2014.

[44] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: an overview and application in radiology," *Insights into imaging*, vol. 9, pp. 611–629, 2018.

[45] J. Hu, J. Lu, and Y.-P. Tan, "Discriminative deep metric learning for face verification in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1875–1882, 2014.

[46] X. Bajrami, B. Gashi, and I. Murturi, "Face recognition performance using linear discriminant analysis and deep neural networks," *International Journal of Applied Pattern Recognition*, vol. 5, no. 3, pp. 240–250, 2018.

[47] P. Norstrom and A. Consulting, "Has covid increased public faith in facial recognition?," *Biometric Technology Today*, vol. 2021, no. 11-12, pp. 5–8, 2021.

[48] A. Anjos and S. Marcel, "Counter-measures to photo attacks in face recognition: a public database and a baseline," in *2011 international joint conference on Biometrics (IJCB)*, pp. 1–7, IEEE, 2011.

[49] S. Marcel, M. S. Nixon, J. Fierrez, and N. Evans, *Handbook of biometric anti-spoofing: Presentation attack detection*, vol. 2. Springer, 2019.

[50] R. Ramachandra and C. Busch, "Presentation attack detection methods for face recognition systems: A comprehensive survey," *ACM Computing Surveys (CSUR)*, vol. 50, no. 1, pp. 1–37, 2017.

[51] S. Jia, G. Guo, and Z. Xu, "A survey on 3d mask presentation attack detection and countermeasures," *Pattern recognition*, vol. 98, p. 107032, 2020.

[52] L. Omar and I. Ivrissimtzis, "Designing a facial spoofing database for processed image attacks," 2016.

[53] L. Omar and I. Ivrissimtzis, "Resilience of luminance based liveness tests under attacks with processed imposter images," 2016.

[54] K. Kollreider, H. Fronthaler, M. I. Faraj, and J. Bigun, "Real-time face detection and motion analysis with application in "liveness" assessment," *IEEE Transactions on Information Forensics and Security*, vol. 2, no. 3, pp. 548–558, 2007.

[55] R. A. Varghese and J. S. Mathew, "Face anti-spoofing methods," *IJSTE-International Journal of Science Technology & Engineering*, vol. 2, no. 4, pp. 318–320, 2015.

[56] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[57] J. Yang, Z. Lei, and S. Z. Li, "Learn convolutional neural network for face anti-spoofing," 2014.

[58] D. Wen, H. Han, and A. K. Jain, "Face spoof detection with image distortion analysis," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 4, pp. 746–761, 2015.

[59] C. Nagpal and S. R. Dubey, "A performance evaluation of convolutional neural networks for face anti spoofing," in *International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, IEEE, 2019.

[60] X. Yang, W. Luo, L. Bao, Y. Gao, D. Gong, S. Zheng, Z. Li, and W. Liu, "Face anti-spoofing: Model matters, so does data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3507–3516, 2019.

[61] K.-Y. Zhang, T. Yao, J. Zhang, Y. Tai, S. Ding, J. Li, F. Huang, H. Song, and L. Ma, "Face anti-spoofing via disentangled representation learning," in *European Conference on Computer Vision*, pp. 641–657, Springer, 2020.

[62] Z. Yu, C. Zhao, Z. Wang, Y. Qin, Z. Su, X. Li, F. Zhou, and G. Zhao, "Searching central difference convolutional networks for face anti-spoofing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[63] M. Asim, Z. Ming, and M. Y. Javed, "Cnn based spatio-temporal feature extraction for face anti-spoofing," in *2017 2nd International Conference on Image, Vision and Computing (ICIVC)*, pp. 234–238, IEEE, 2017.

[64] X. Song, X. Zhao, L. Fang, and T. Lin, "Discriminative representation combinations for accurate face spoofing detection," *Pattern Recognition*, vol. 85, pp. 220–231, 2019.

[65] Y. A. U. Rehman, L.-M. Po, and J. Komulainen, "Enhancing deep discriminative feature maps via perturbation for face presentation attack detection," *Image Vis. Comput.*, vol. 94, p. 103858, 2020.

[66] M. Khammari, "Robust face anti-spoofing using cnn with lbp and wld," *IET Image Processing*, vol. 13, no. 11, pp. 1880–1884, 2019.

[67] A. George, Z. Mostaani, D. Geissenbuhler, O. Nikisins, A. Anjos, and S. Marcel, "Biometric face presentation attack detection with multi-channel convolutional neural network," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 42–55, 2019.

[68] Z. Zhang, J. Yan, S. Liu, Z. Lei, D. Yi, and S. Z. Li, "A face antispoofing database with diverse attacks," in *2012 5th IAPR International Conference on Biometrics (ICB)*, pp. 26–31, 2012.

[69] Y. Jia, J. Zhang, S. Shan, and X. Chen, "Unified unsupervised and semi-supervised domain adaptation network for cross-scenario face anti-spoofing," *Pattern Recognition*, vol. 115, p. 107888, 2021.

[70] Y. Liu, J. Stehouwer, A. Jourabloo, and X. Liu, "Deep tree learning for zero-shot face anti-spoofing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[71] Y. Qin, C. Zhao, X. Zhu, Z. Wang, Z. Yu, T. Fu, F. Zhou, J. Shi, and Z. Lei, "Learning meta model for zero- and few-shot face anti-spoofing," in *AAAI*, 2020.

[72] Y. A. U. Rehman, L.-M. Po, and M. Liu, "Slnet: Stereo face liveness detection via dynamic disparity-maps and convolutional neural network," *Expert Systems with Applications*, vol. 142, p. 113002, 2020.

[73] B. Wu, M. Pan, and Y. Zhang, *A Review of Face Anti-spoofing and Its Applications in China*, pp. 35–43. 01 2020.

[74] J. Connell, N. Ratha, J. Gentile, and R. Bolle, "Fake iris detection using structured light," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 8692–8696, IEEE, 2013.

[75] T. Kim and Y. Kim, "Suppressing spoof-irrelevant factors for domain-agnostic face anti-spoofing," *IEEE Access*, vol. 9, pp. 86966–86974, 2021.

[76] R. Ramachandra and C. Busch, "Presentation attack detection methods for face recognition systems: A comprehensive survey," *ACM Computing Surveys (CSUR)*, vol. 50, no. 1, pp. 1–37, 2017.

[77] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[78] A. Costa-Pazo, S. Bhattacharjee, E. Vazquez-Fernandez, and S. Marcel, "The replay-mobile face presentation-attack database," in *2016 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pp. 1–7, 2016.

[79] Z. Boulkenafet, J. Komulainen, L. Li, X. Feng, and A. Hadid, "Oulu-npu: A mobile face presentation attack database with real-world variations," in *2017 12th IEEE international conference on automatic face & gesture recognition (FG 2017)*, pp. 612–618, IEEE, 2017.

[80] H. Li, W. Li, H. Cao, S. Wang, F. Huang, and A. C. Kot, "Unsupervised domain adaptation for face anti-spoofing," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 7, pp. 1794–1809, 2018.

[81] H. Li, W. Li, H. Cao, S. Wang, F. Huang, and A. C. Kot, "Unsupervised domain adaptation for face anti-spoofing," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 7, pp. 1794–1809, 2018.

[82] N. Erdogmus and S. Marcel, "Spoofing in 2d face recognition with 3d masks and anti-spoofing with kinect," in *2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pp. 1–6, 2013.

[83] D. Yi, "Learning face representation from scratch," *CoRR*, vol. 1411, p. 7923, 2014.

[84] M. Grgic, K. Delac, and S. Grgic, "Scface–surveillance cameras face database," *Multimedia tools and applications*, vol. 51, no. 3, pp. 863–879, 2011.

[85] M. Wang, W. Deng, J. Hu, X. Tao, and Y. Huang, "Racial faces in the wild: Reducing racial bias by information maximization adaptation network," pp. 692–702, 10 2019.

[86] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," in *ECCV*, 2016.

[87] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*, vol. 1, pp. I–I, IEEE, 2001.

[88] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[89] S. Lim, Y. Gwak, W. Kim, J.-H. Roh, and S. Cho, "One-class learning method based on live correlation loss for face anti-spoofing," *IEEE Access*, vol. 8, pp. 201635–201648, 2020.

[90] S. Fatemifar, M. Awais, A. Akbari, and J. Kittler, "Particle swarm and pattern search optimisation of an ensemble of face anomaly detectors," in *2021 IEEE International Conference on Image Processing (ICIP)*, pp. 3622–3626, IEEE, 2021.

[91] Z. Yu, C. Zhao, Z. Wang, Y. Qin, Z. Su, X. Li, F. Zhou, and G. Zhao, "Searching central difference convolutional networks for face anti-spoofing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5295–5305, 2020.

[92] S. Fatemifar, M. Awais, S. R. Arashloo, and J. Kittler, "Combining multiple one-class classifiers for anomaly based face spoofing attack detection," in *2019 International Conference on Biometrics (ICB)*, pp. 1–7, IEEE, 2019.

[93] Y. Yao, G. L. Marcialis, M. Pontil, P. Frasconi, and F. Roli, "Combining flat and structured representations for fingerprint classification with recursive neural networks and support vector machines," *Pattern Recognit.*, vol. 36, pp. 397–406, 2003.

[94] J. Daugman, "New methods in iris recognition," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 37, no. 5, pp. 1167–1175, 2007.

[95] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM computing surveys (CSUR)*, vol. 41, no. 3, pp. 1–58, 2009.

[96] C. C. Aggarwal, "An introduction to outlier analysis," in *Outlier analysis*, pp. 1–34, Springer, 2017.

[97] P. Perera, P. Oza, and V. M. Patel, "One-class classification: A survey," *arXiv preprint arXiv:2101.03064*, 2021.

[98] M. Sabokrou, M. Khalooei, M. Fathy, and E. Adeli, "Adversarially learned one-class classifier for novelty detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[99] S. Rahimzadeh Arashloo, "Unknown face presentation attack detection via localised learning of multiple kernels," 2022.

[100] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural computation*, vol. 13, no. 7, pp. 1443–1471, 2001.

[101] J. Kittler, W. Christmas, T. De Campos, D. Windridge, F. Yan, J. Illingworth, and M. Osman, "Domain anomaly detection in machine perception: A system architecture and taxonomy," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 5, pp. 845–859, 2013.

[102] O. Nikisins, A. Mohammadi, A. Anjos, and S. Marcel, "On effectiveness of anomaly detection approaches against unseen presentation attacks in face anti-spoofing," in *2018 International Conference on Biometrics (ICB)*, pp. 75–81, IEEE, 2018.

[103] J. Galbally, S. Marcel, and J. Fierrez, "Image quality assessment for fake biometric detection: Application to iris, fingerprint, and face recognition," *IEEE transactions on image processing*, vol. 23, no. 2, pp. 710–724, 2013.

[104] D. Wen, H. Han, and A. K. Jain, "Face spoof detection with image distortion analysis," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 4, pp. 746–761, 2015.

[105] A. Costa-Pazo, S. Bhattacharjee, E. Vazquez-Fernandez, and S. Marcel, "The replay-mobile face presentation-attack database," in *2016 international conference of the Biometrics Special Interest Group (BIOSIG)*, pp. 1–7, IEEE, 2016.

[106] S. Fatemifar, S. R. Arashloo, M. Awais, and J. Kittler, "Spoofing attack detection by anomaly detection," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8464–8468, IEEE, 2019.

[107] S. Fatemifar, M. Awais, A. Akbari, and J. Kittler, "A stacking ensemble for anomaly based client-specific face spoofing detection," in *2020 IEEE International Conference on Image Processing (ICIP)*, pp. 1371–1375, IEEE, 2020.

[108] M. Ibsen, L. J. González-Soler, C. Rathgeb, P. Drozdowski, M. Gomez-Barrero, and C. Busch, "Differential anomaly detection for facial images," in *2021 IEEE International Workshop on Information Forensics and Security (WIFS)*, pp. 1–6, IEEE, 2021.

[109] P. Oza and V. M. Patel, "One-class convolutional neural network," *IEEE Signal Processing Letters*, vol. 26, no. 2, pp. 277–281, 2018.

[110] I. Chingovska and A. R. dos Anjos, "On the use of client identity information for face antispoofing," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 4, pp. 787–796, 2015.

[111] V. Štruc and N. Pavešic, "Photometric normalization techniques for illumination invariance," in *Advances in face image analysis: Techniques and technologies*, pp. 279–300, IGI Global, 2011.

[112] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

[113] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[114] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations (ICLR 2015)*, Computational and Biological Learning Society, 2015.

[115] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," 2015.

[116] D. H. Wolpert, "Stacked generalization," *Neural networks*, vol. 5, no. 2, pp. 241–259, 1992.

[117] S. R. Arashloo, "Unseen face presentation attack detection using sparse multiple kernel fisher null-space," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 31, no. 10, pp. 4084–4095, 2020.

[118] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of ICNN'95-international conference on neural networks*, vol. 4, pp. 1942–1948, IEEE, 1995.

[119] R. Hooke and T. A. Jeeves, ""direct search"solution of numerical and statistical problems," *Journal of the ACM (JACM)*, vol. 8, no. 2, pp. 212–229, 1961.

[120] S. Fatemifar, S. Asadi, M. Awais, A. Akbari, and J. Kittler, "Face spoofing detection ensemble via multistage optimisation and pruning," *Pattern Recognition Letters*, vol. 158, pp. 1–8, 2022.

[121] Y. Zhang, M. Zhao, L. Yan, T. Gao, and J. Chen, "Cnn-based anomaly detection for face presentation attack detection with multi-channel images," in *2020 IEEE International Conference on Visual Communications and Image Processing (VCIP)*, pp. 189–192, IEEE, 2020.

[122] S. R. Arshloo, "Matrix-regularized one-class multiple kernel learning for unseen face presentation attack detection," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 4635–4647, 2021.

[123] X. Huang, J. Xia, and L. Shen, "One-class face anti-spoofing based on attention auto-encoder," (Berlin, Heidelberg), Springer-Verlag, 2021.

[124] Y. Liu*, A. Jourabloo*, and X. Liu, "Learning deep models for face anti-spoofing: Binary or auxiliary supervision," in *In Proceeding of IEEE Computer Vision and Pattern Recognition*, (Salt Lake City, UT), June 2018.

[125] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[126] J. Galbally, S. Marcel, and J. Fierrez, "Biometric antispoofing methods: A survey in face recognition," *IEEE Access*, vol. 2, pp. 1530–1552, 2014.

[127] L. Li, P. L. Correia, and A. Hadid, "Face recognition under spoofing attacks: countermeasures and research directions," *Iet Biometrics*, vol. 7, no. 1, pp. 3–14, 2018.

[128] L. Omar and I. Ivrissimtzis, "Designing a facial spoofing database for processed image attacks," 2016.

[129] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *stat*, vol. 1050, p. 4, 2019.

[130] A. E. Cinà, A. Torcinovich, and M. Pelillo, "A black-box adversarial attack for poisoning clustering," *Pattern Recognition*, vol. 122, p. 108306, 2022.

[131] N. Narodytska and S. P. Kasiviswanathan, "Simple black-box adversarial perturbations for deep networks," *arXiv preprint arXiv:1612.06299*, 2016.

[132] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: a simple and accurate method to fool deep neural networks," in *CVPR*, pp. 2574–2582, 2016.

[133] E. Sarkar, P. Korshunov, L. Colbois, and S. Marcel, "Are gan-based morphs threatening face recognition?," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2959–2963, IEEE, 2022.

[134] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition," in *Proceedings of the 2016 acm sigsac conference on computer and communications security*, pp. 1528–1540, 2016.

[135] B. Zhang, B. Tondi, and M. Barni, "Adversarial examples for replay attacks against cnn-based face recognition with anti-spoofing capability," *Computer Vision and Image Understanding*, vol. 197-198, p. 102988, 2020.

[136] T. Ayanwola, A. Oludele, and M. Agbaje, "Enhancing face spoofing attack detection: Performance evaluation of a vgg-19 cnn model," *Acadlore Trans. Mach. Learn*, vol. 2, no. 2, pp. 84–98, 2023.

[137] O. Lucena, A. Junior, V. Moia, R. Souza, E. Valle, and R. Lotufo, "Transfer learning using convolutional neural networks for face anti-spoofing," in *Image Analysis and Recognition: 14th International Conference, ICIAR 2017, Montreal, QC, Canada, July 5–7, 2017, Proceedings 14*, pp. 27–34, Springer, 2017.

[138] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al.*, "Imagenet large scale visual recognition challenge," *IJCV*, vol. 115, no. 3, pp. 211–252, 2015.

[139] H. H. Tan and K. H. Lim, "Vanishing gradient mitigation with deep learning neural network optimization," in *2019 7th International Conference on Smart Computing & Communications (ICSCC)*, pp. 1–4, 2019.

[140] R. Lukac and K. N. Plataniotis, *Color image processing: methods and applications*. CRC press, 2018.

[141] I. Chingovska, A. Anjos, and S. Marcel, "On the effectiveness of local binary patterns in face anti-spoofing," in *BIOSIG)*, pp. 1–7, IEEE, 2012.

[142] Y. Chen, T. Wang, J. Wang, P. Shi, G. Shan, and H. Snoussi, "Towards good practices in face anti-spoofing: An image reconstruction based method," in *2019 Chinese Automation Congress (CAC)*, pp. 4700–4705, 2019.

[143] F. Xiong and W. AbdAlmageed, "Unknown presentation attack detection with face rgb images," in *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pp. 1–9, IEEE, 2018.

[144] Y. Xia, X. Cao, F. Wen, G. Hua, and J. Sun, "Learning discriminative reconstructions for unsupervised outlier removal," in *Proc. ICCV*, pp. 1511–1519, IEEE), 2015.

[145] D. Yang, J. Lai, and L. Mei, "Deep representations based on sparse auto-encoder networks for face spoofing detection," in *Chinese Conference on Biometric Recognition*, pp. 620–627, Springer, 2016.

[146] A. George and S. Marcel, "On the effectiveness of vision transformers for zero-shot face anti-spoofing," in *Proc. IJCB*, pp. 1–8, IEEE, 2021.

[147] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[148] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel, "Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness," 2018.

[149] Z. Wang, Q. Wang, W. Deng, and G. Guo, "Face anti-spoofing using transformers with relation-aware mechanism," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 4, no. 3, pp. 439–450, 2022.

[150] H.-P. Huang, D. Sun, Y. Liu, W.-S. Chu, T. Xiao, J. Yuan, H. Adam, and M.-H. Yang, "Adaptive transformers for robust few-shot cross-domain face anti-spoofing," in *European Conference on Computer Vision*, pp. 37–54, Springer, 2022.

[151] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.

[152] R. Liu, W. Liu, Z. Zheng, L. Wang, L. Mao, Q. Qiu, and G. Ling, "Anomaly-gan: A data augmentation method for train surface anomaly detection," *Expert Systems with Applications*, p. 120284, 2023.

[153] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *2008 eighth ieee international conference on data mining*, pp. 413–422, IEEE, 2008.

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

[155] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International conference on machine learning*, pp. 10347–10357, PMLR, 2021.

[156] I. Chingovska, A. Anjos, and S. Marcel, "On the effectiveness of local binary patterns in face anti-spoofing," in *The International Conference of Biometrics Special Interest Group (BIOSIG)*, pp. 1–7, IEEE, 2012.

[157] I. Standard, "Information technology—biometric presentation attack detection—part 1: Framework," *ISO: Geneva, Switzerland*, vol. 6, 2016.

[158] F. Xiong and W. AbdAlmageed, "Unknown Presentation Attack Detection with Face RGB Images," in *Proc. BTAS*, pp. 1–9, 2018.

[159] "ID R&D whitepaper: Mitigating Demographic Bias in Facial Presentation Attack Detection," 2022.

[160] A. Van Den Oord, O. Vinyals, *et al.*, "Neural discrete representation learning," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[161] J. A. Hartigan and P. M. Hartigan, "The dip test of unimodality," *The annals of Statistics*, pp. 70–84, 1985.

[162] D. Jiménez-Cabello and D. Pérez-Cabo, "Deep anomaly detection for generalized face anti-spoofing," in *Actas del IV Machine Learning Workshop*, pp. 1–31, University of A Coruña, 2019.

[163] O. Nikisins, A. George, and S. Marcel, "Domain adaptation in multi-channel autoencoder based features for robust face anti-spoofing," in *2019 International Conference on Biometrics (ICB)*, pp. 1–8, 2019.

[164] Y. Zhang, M. Zhao, L. Yan, T. Gao, and J. Chen, "Cnn-based anomaly detection for face presentation attack detection with multi-channel images," *IEEE International Conference on Visual Communications and Image Processing (VCIP)*, pp. 189–192, 2020.

[165] B. S. F. Silva and M. Costa-Abreu, "Exploring bias analysis on judicial data using machine learning techniques," in *2022 12th International Conference on Pattern Recognition Systems (ICPRS)*, pp. 1–7, 2022.

[166] A. Yapo and J. Weiss, "Ethical implications of bias in machine learning," 2018.

[167] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021.

[168] H. Suresh and J. Guttag, "A framework for understanding unintended consequences of machine learning," p. 8, 2019.

[169] A. Olteanu, C. Castillo, F. Diaz, and E. Kıcıman, "Social data: Biases, methodological pitfalls, and ethical boundaries," *Frontiers in Big Data*, vol. 2, p. 13, 2019.

[170] R. Singh, A. Agarwal, M. Singh, S. Nagpal, and M. Vatsa, "On the robustness of face recognition algorithms against attacks and bias," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 13583–13589, 2020.

[171] P. J. Phillips, F. Jiang, A. Narvekar, J. Ayyad, and A. J. O'Toole, "An other-race effect for face recognition algorithms," *ACM Trans. Appl. Percept.*, vol. 8, feb 2011.

[172] R. V. Garcia, L. Wandzik, L. Grabner, and J. Krueger, "The harms of demographic bias in deep face recognition research," in *2019 International Conference on Biometrics (ICB)*, pp. 1–6, IEEE, 2019.

[173] S. Glüge, M. Amirian, D. Flumini, and T. Stadelmann, "How (not) to measure bias in face recognition networks," in *IAPR Workshop on Artificial Neural Networks in Pattern Recognition*, pp. 125–137, Springer, 2020.

[174] I. Serna, A. Peña, A. Morales, and J. Fierrez, "Insidebias: Measuring bias in deep networks and application to face gender biometrics," in *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 3720–3727, IEEE, 2021.

[175] J. G. Cavazos, P. J. Phillips, C. D. Castillo, and A. J. O'Toole, "Accuracy comparison across face recognition algorithms: Where are we on measuring race bias?," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2020.

[176] T. de Freitas Pereira and S. Marcel, "Fairness in biometrics: a figure of merit to assess biometric verification systems," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 4, no. 1, pp. 19–29, 2021.

[177] A. Liu, X. Li, J. Wan, S. Escalera, H. J. Escalante, M. Madadi, Y. Jin, Z. Wu, X. Yu, Z. Tan, Q. Yuan, R. Yang, B. Zhou, G. Guo, and S. Z. Li, "Cross-ethnicity face anti-spoofing recognition challenge: A review," 2020.

[178] N. Alshareef, X. Yuan, K. Roy, and M. Atay, "A study of gender bias in face presentation attack and its mitigation," *Future Internet*, vol. 13, no. 9, p. 234, 2021.

[179] A. Amini, A. P. Soleimany, W. Schwarting, S. N. Bhatia, and D. Rus, "Uncovering and mitigating algorithmic bias through learned latent structure," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 289–295, 2019.

[180] M. Fang, N. Damer, F. Kirchbuchner, and A. Kuijper, "Demographic bias in presentation attack detection of iris recognition systems," in *2020 28th European Signal Processing Conference (EUSIPCO)*, pp. 835–839, IEEE, 2021.

[181] M. Wang, W. Deng, J. Hu, X. Tao, and Y. Huang, "Racial faces in the wild: Reducing racial bias by information maximization adaptation network," pp. 692–702, 10 2019.

[182] A. Liu, Z. Tan, J. Wan, S. Escalera, G. Guo, and S. Z. Li, "Casia-surf cefa: A benchmark for multi-modal cross-ethnicity face anti-spoofing," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1179–1187, January 2021.

[183] S. N. Marimont and G. Tarroni, "Anomaly detection through latent space restoration using vector quantized variational autoencoders," in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pp. 1764–1767, IEEE, 2021.

[184] L. Wang, D. Zhang, J. Guo, and Y. Han, "Image anomaly detection using normal data only by latent space resampling," *Applied Sciences*, vol. 10, no. 23, p. 8660, 2020.

[185] M. Hearst, S. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intelligent Systems and their Applications*, vol. 13, no. 4, pp. 18–28, 1998.

[186] R. L. Plackett, "Karl pearson and the chi-squared test," *International Statistical Review / Revue Internationale de Statistique*, vol. 51, no. 1, pp. 59–72, 1983.

[187] H. B. Mann and D. R. Whitney, "On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other," *The Annals of Mathematical Statistics*, vol. 18, no. 1, pp. 50 – 60, 1947.

[188] H. Liu, Z. Zhao, X. Chen, R. Yu, and Q. She, "Using the vq-vae to improve the recognition of abnormalities in short-duration 12-lead electrocardiogram records," *Computer Methods and Programs in Biomedicine*, vol. 196, p. 105639, 2020.

[189] H. Zenati, C.-S. Foo, B. Lecouat, G. Manek, and V. Chandrasekhar, "Efficient gan-based anomaly detection," *ArXiv*, vol. abs/1802.06222, 2018.

[190] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, "Unsupervised anomaly detection with generative adversarial networks to guide marker discovery," in *International conference on information processing in medical imaging*, pp. 146–157, Springer, 2017.

[191] V. Gupta, M. Nishigaki, and T. Ohki, "Unsupervised biometric anti-spoofing using generative adversarial networks," *International Journal of Informatics Society*, vol. 11, no. 1, 2019.