Boosting the Generalizability and Fairness of Presentation Attack Detection

Zur Erlangung des akademischen Grades Doktor-Ingenieur (Dr.-Ing.)
Genehmigte Dissertation von Meiling Fang aus China
Tag der Einreichung: 09. March 2023, Tag der Prüfung: 21. April 2023

1. Gutachten: Prof.Dr. Arjan Kuijper
2. Gutachten: Prof.Dr. Dieter W. Fellner
3. Gutachten: Prof.Dr. Julian Fierrez

Darmstadt, Technische Universität Darmstadt

TECHNISCHE
UNIVERSITÄT
DARMSTADT

Computer Science
Department

Interactive Graphics
Systems Group

Boosting the Generalizability and Fairness of Presentation Attack Detection

Accepted doctoral thesis by Meiling Fang

Date of submission: 09. March 2023
Date of thesis defense: 21. April 2023

Darmstadt, Technische Universität Darmstadt

Erklärungen laut Promotionsordnung

§ 8 Abs. 1 lit. c PromO

Ich versichere hiermit, dass die elektronische Version meiner Dissertation mit der schriftlichen Version übereinstimmt.

§ 8 Abs. 1 lit. d PromO

Ich versichere hiermit, dass zu einem vorherigen Zeitpunkt noch keine Promotion versucht wurde. In diesem Fall sind nähere Angaben über Zeitpunkt, Hochschule, Dissertationsthema und Ergebnis dieses Versuchs mitzuteilen.

§ 9 Abs. 1 PromO

Ich versichere hiermit, dass die vorliegende Dissertation selbstständig und nur unter Verwendung der angegebenen Quellen verfasst wurde.

§ 9 Abs. 2 PromO

Die Arbeit hat bisher noch nicht zu Prüfungszwecken gedient.

Darmstadt, 09. March 2023

Meiling Fang

Abstract

The vulnerability of biometric recognition to presentation attacks (PAs) has been widely recognized and has attracted increasing attention as it enables attackers to impersonate authentic users. Presentation attack detection (PAD), aiming at automatically catching PAs, is an essential technology to secure biometric systems from PAs such as printed photos and replayed videos. Despite the considerable exploration and remarkable progress in PAD performance, two major issues still constitute a gap in technology. The first is the lack of proper understanding of the fairness of such algorithms over human-related attributes, and the second is the low performance generalizability over variabilities such as unknown attack types and capture environments. These challenges drive the main contributions of this thesis towards analyzing and boosting the fairness and generalizability of PAD.

PAD fairness over different human attributes is extremely understudied. Such under-exploration is mainly due to the lack of suitable data. Towards enabling the fairness assessment and enhancement in face PAD, this thesis first introduces a combined attribute annotated PAD dataset, including both demographic and non-demographic attribute labels. Meanwhile, this thesis presents a new metric, accuracy balanced fairness, to simultaneously represent both the PAD fairness and the absolute PAD performance. Then, a comprehensive analysis of fairness in face PAD is conducted to study its relation to the nature of training data and the methodology of decision threshold selection. Guided by the outcomes of these analyses, a data augmentation method, namely FairSWAP, is successfully proposed to enhance the fairness of face PAD.

In addition to the PAD generalizability over human-related attributes, seen as fairness, another emerging challenge that encountered face PAD during the COVID-19 pandemic is the PAD generalizability to subjects wearing facial masks. To address this issue, this thesis first provides a collaborative real mask attack dataset involving the conventional unmasked bona fide and attacks, masked bona fide sample, novel attacks with faces wearing masks, and attacks with real masks placed on spoof faces. This thesis performs a set of extensive experiments to investigate the impact of masked faces on recognition vulnerability and PAD behaviour. Observing the degradation of PAD performance caused by the facial masks, this thesis presents a solution to target this issue by refining the partial attack supervision and the regional weighted inference.

The third part of this thesis targets the more conventional PAD generalizability issues, such as variabilities in attack creation and capture scenarios. Aiming to boost the generalizability of face PAD, this thesis proposes to leverage the information from the frequency domain in an optimized manner, assisting the information in the spatial domain to learn a more generalized representation under intra-dataset and cross-dataset settings. With a focus on enhancing the generalizability of iris PAD, this thesis proposes a micro-stripe analyses solution that leverages the benefit of the spatially aware processing of well-defined regions in the iris and its border with the sclera. This thesis further introduces a novel attention-based deep pixel-wise binary supervision method, A-PBS, for iris PAD. This solution aims to capture the fine-grained pixel/patch-level attack clues and automatically locate regions that contribute the most to an accurate PAD decision. The generalizability of the proposed iris PAD solutions is demonstrated under real-world cross-testing cases, including cross-attack, cross-dataset, and cross-spectrum settings.

To summarize, this thesis first provides a much-needed comprehensive analysis of fairness in PAD, leading to a well-founded and integrable fairness enhancement solution. Then, it presents detailed investigations of the masked face PAD challenge along with a technical solution towards improving the masked face PAD performance. The thesis then presents a set of novel contributions to boost the generalizability of face and iris PAD techniques. This thesis thus yields practically-aware advancements in understanding and mitigating vulnerabilities of biometric systems and lays the groundwork for future research into developing and deploying generalized PAD systems.

Zusammenfassung

Die Anfälligkeit von biometrischer Erkennung für Presentation Attacks (PAs) ist weithin bekannt und hat zunehmend Aufmerksamkeit erregt, da sie Angreifern ermöglicht, sich als authentische Benutzer auszugeben. Die automatisierte Erkennung von Presentation Attacks (Presentation Attack Detection (PAD)) ist eine wesentliche Technologie um biometrische Systeme vor solchen Angriffen, etwa mit gedruckten Fotos oder abgespielten Videos, zu schützen. Trotz des beachtlichen und bemerkenswerten Fortschritts bezüglich der Erkennungsraten im Bereich PAD zeigen sich noch zwei Hauptprobleme. Das erste Problem ist die mangelnde Kenntnis bezüglich der Fairness der Algorithmen in Bezug auf menschenbezogene Attribute und das zweite Problem ist die geringe Generalisierbarkeit der Algorithmen in Bezug auf Veränderungen wie etwa unbekannte Angriffstypen und unbekannte Aufnahmeumgebungen. Die Herausforderungen, die sowohl aus der Analyse, der Steigerung der Fairness sowie der Generalisierbarkeit von PAD-Algorithmen bestehen, motivieren den Kern dieser Dissertation.

Die Fairness von PAD-Algorithmen bezüglich verschiedener personenbezogener Attribute ist noch sehr wenig erforscht, was vor allem auf den Mangel an geeigneten Daten zurückzuführen ist. Um die Fairness von PAD-Algorithmen für Gesichtserkennungssysteme zu bewerten und diese auch zu erhöhen, wird in dieser Thesis zuerst ein kombinierter PAD-Datensatz eingeführt, welcher mit demografischen sowie nicht-demografischen Attributen annotiert ist. Des Weiteren stellt diese Arbeit eine neue Metrik, acurracy balanced fairness, vor, um gleichzeitig die Fairness und die Erkennungsgenauigkeit von PAD-Algorithmen zu messen. Anschließend wird eine umfassende Analyse der Fairness dieser Algorithmen für Gesichter durchgeführt. Dabei werden die Zusammenhänge zwischen Fairness und der Art der Trainingsdaten sowie zwischen Fairness und der Auswahlmethodik der Entscheidungsgrenzen untersucht. Angeleitet von den Ergebnissen dieser Analysen wird eine neue Datenaugmentierungsmethode, FairSWAP, vorgeschlagen, um die Fairness von PAD-Algorithmen für Gesichter zu verbessern.

Neben der Generalisierbarkeit von PAD-Algorithmen bezüglich personenbezogenen Attributen, die als Fairness angesehen werden kann, ist die Generalisierbarkeit von PAD-Algorithmen auf Personen, die Masken tragen, eine Herausforderung, welche durch die COVID-19 Pandemie aufgekommen ist. Um dieses Problem anzugehen wird in dieser

Arbeit zuerst ein Datensatz vorgestellt, welcher authentische maskierte und unmaskierte Gesichter, sowie Angriffe mit Gesichtern, die Masken tragen und Angriffe mit echten Masken auf nicht authentischen Gesichtern, enthält. In dieser Thesis werden umfangreiche Experimente durchgeführt, um die Auswirkungen von maskierten Gesichtern auf die Erkennungsanfälligkeit und die Genauigkeit der PAD-Algorithmen zu untersuchen. Angesichts der Verschlechterung der PAD-Algorithmen bezüglich ihrer Genauigkeit verursacht durch die Masken, wird in dieser Arbeit eine Lösung vorgestellt, die auf der Verfeinerung der partial attack supervision und der regional weighted inference basiert.

Der dritte Teil dieser Arbeit befasst sich mit den konventionellen PAD Generalisierungsproblemen, wie z.B. Unterschiede bei der Erstellung von Angriffen und unterschiedliche biometrische Aufnahmeszenarien. Mit dem Ziel, die Generalisierung von PAD-Algorithmen für Gesichter zu verbessern, schlägt diese Arbeit vor, zusätzlich zum räumlichen Bereich, Informationen aus dem Frequenzbereich in einer optimierten Form zu nutzen, um eine generalisierte Repräsentation zu lernen, sowohl für einzelne Datenbanken, als auch datenbankübergreifend. Um die Generalisierbarkeit von PAD-Algorithmen für Iriserkennung zu verbessern wird in dieser Arbeit eine Lösung vorgeschlagen, die auf der Analyse von micro-stripes basiert und den Vorteil der spatialen Verarbeitung von klar definierten Regionen der Iris und ihrer Abgrenzung mit der Sklera nutzt. Diese Dissertation führt zusätzlich eine neue Methodik ein, A-PBS, welche eine attention-based deep pixel-wise binary supervision Methode für Iris-PAD ist. Diese Methode zielt darauf ab, subtile Angriffshinweise auf Pixel- bzw. Patch-Ebene zu erfassen und automatisch die Regionen zu finden, welche am meisten zu einer richtigen PAD-Entscheidung beitragen. Die Generalisierbarkeit der vorgeschlagenen Iris-PAD-Lösungen wird unter realen Cross-Testing-Fällen demonstriert, einschließlich angriffsübergreifender, datensatzübergreifender und spektrenübergreifender Einstellungen.

Zusammenfassend bietet diese Arbeit zunächst eine notwendige, umfassende Analyse der Fairness von PAD-Algorithmen, die eine integrierbare Lösung zur Verbesserung der Fairness motiviert. Anschließend werden detaillierte Untersuchungen zur Herausforderung von Gesichtern, welche Masken tragen für PAD-Algorithmen präsentiert und zusammen mit einer technischen Lösung zur Verbesserung der PAD-Erkennungsraten von maskierten Gesichtern vorgestellt. Dann wird eine Reihe von neuartigen Lösungen präsentiert, welche die Generalisierbarkeit von PAD-Algorithmen für Gesicht- und Iriserkennung verbessert. Diese Arbeit liefert somit praxisnahe Fortschritte bei der Reduzierung und dem Verständnis von Schwachstellen von biometrischen Systemen und legt den Grundstein für die zukünftige Forschung zur Entwicklung und zum Einsatz generalisierter PAD-Erkennungssysteme.

Acknowledgement

First and foremost I would like to thank my supervisor, Prof.Dr. Arjan Kuijper, for his invaluable guidance, encouragement, and support throughout my this work. I would like to express my sincere gratitude to Prof. Dr. techn. Dieter W. Fellner for co-refereeing this work. Special thanks go to Prof.Dr. Julian Fierrze for acting as co-referee.

The deepest and sincerest gratitude extends to my supervisor and friend, Dr. Naser Damer, for his continuous and invaluable guidance throughout my research. I would like to thank him for his insightful feedback, motivation, support, and a great sense of humor. I consider myself one of the luckiest people who benefited from his knowledge and character.

I would also like to acknowledge all my friends and colleagues at the Competence Center for Smart Living & Biometric Technology (SLBT) of the Fraunhofer IGD. A Special thank goes to the head of the SLBT department, Florian Kirchbuchner, for providing an excellent academic environment and constant support, which greatly contributed to the successful completion of my doctoral program. The special and deepest thank extends to my friend and colleague, Dr. Fadi Boutros, working with him is definitely one of the most joyful experience. Deepest thanks also go to Naser, Fadi, Marco, and Jan for their outstanding efforts in reviewing this thesis.

My greatest thanks go to my students, whom I have learned from the most. Thank you, Hamza, Wufei, Nour, Lu, for your excellent work.

I would like to extend my appreciation to all my friends. I would like to give special thanks to Xin for her positive energy and contagious enthusiasm, which never failed to lift my spirits. The deepest thank goes to Yang for being my constant source of motivation and emotional support. Your unwavering belief in me kept me going even when the going got tough. I am deeply indebted to my parents for their unconditional love, encouragement, and moral support throughout my life.

To everyone who played a role in my PhD journey, thank you.

Contents

Glossary

ABC Automated Border Control.

ABF Accuracy Balanced Fairness.

ACER Average Classification Error Rate.

APCER Attack Presentation Classification Error Rate.

BCE Binary Cross-Entropy.

BPCER Bona fide Presentation Classification Error Rate.

BSIF Binarized Statistical Image Features.

CAAD-PAD Combined Attribute Annotated Presentation Attack Detection Dataset.

CBAM Convolutional Block Attention Module.

CCR Correct Classification Rate.

CE Cross-Entropy.

CLAHE Contrast Limited Adaptive Histogram Equalization.

CNN Convolutional Neural Network.

CRMA Collaborative Real Mask Attack Database.

D-EER Detection Equal Error Rate.

DET Detection Error Tradeoff.

EER Equal Error Rate.

FC Fully-connected.

FDR Fairness Discrepancy Rate.

FMR False Match Rate.

FNMR False Non-match Rate.

FR Face Recognition.

GRU Gated Recurrent Unit.

HAM Hierarchical Attention Mechanism.

HTER Half Total of Error Rate.

IAPMR Imposter Attack Presentation Match Rate.

LBP Local Binary Pattern.

LSTM Long Short-Term Memory.

MFD Multi-level Frequency Decomposition.

MSA Micro Stripes Analyses.

MSE Mean Squard Error.

NAS Neural Architecture Search.

NIR Near-Infrared light.

NIST National Institute of Standards and Technology.

ODTA Operational Decision Threshold Assignment.

PA Presentation Attack.

PAD Presentation Attack Detection.

PAI Presentation Attack Instrument.

PAs Presentation Attacks.

PAVID Presentation Attack Video Iris Database.

PCA Principal Component Analysis.

PIN Personal Identification Number.

PPG PhotoPlethysmoGraphy.

PSF Photometric Stereo Features.

RNN Recurrent Neural Network.

ROC Receiver Operating Characteristic.

SOTA State of The Art.

SSIM Structural Similarity.

SVM Support Vector Machine.

TDR True Detection Rate.

TSM Temporal Shift Module.

1. Introduction

Biometrics refers to the automated recognition of individuals based on their physiological or behavioural characteristics [130]. Biometrics relies on essentially immutable and unique facets of an individual, and thus is more reliable in recognizing identity in comparison with traditional approaches of recognition, such as knowledge-based (e.g., passwords and Personal Identification Number (PIN) codes) and physical-based (e.g., tokens and cards) approaches. In addition, biometric traits cannot be forgotten, lost or stolen, thereby obviating the need to remember passwords or carry tokens. Considering such properties and remarkable accuracies achieved by biometric recognition algorithms, biometrics has been widely deployed in various security scenarios, such as mobile device access [192], automated border control [66], and forensic applications [30].

When deploying biometrics in practical applications, proper biometric characteristics should be selected based on the weighting of several factors, including universality, uniqueness, permanence, collectability, performance, acceptability, and circumvention [11]. Circumvention, describing the ease with which a trait might be imitated using an artifact or substitute, is associated with the vulnerability of biometrics. Numerous studies [14, 43, 168, 169, 233, 261] have shown that biometric recognition is vulnerable to diverse attacks, in which Presentation Attack (PA) is the most widely attempted. PA refers to presenting an attack instrument, such as a high-resolution printed face photo, artificial fingers, or a recorded audio voice, to the biometric sensor with the goal of interfering with the operation of the biometric recognition systems [127].

For instance, early fingerprint-based recognition systems could be compromised by relatively crude Presentation Attacks (PAs). In 2012, the Chaos Computer Club successfully bypassed the biometric security of Apple TouchID deployed in iPhone 5s using a fake fingerprint from a glass surface [31]. A doctor used fake fingers made of silicon to sign in absent colleagues in 2013 [7]. In the case of face trait, a parade of tech commentators, journalists, and reviewers attacked the biometric systems of Samsung Galaxy S10 and S10+ devices to unlock them only with YouTube videos and static images on other devices in 2019 [196]. Also, digital security experts at Bkav released footage showing that they have fooled the face recognition systems on the iPhone X by a 3D mask made of stone powder with 2D images of the user's eyes in 2017 [220]. Later, Tencent researchers bypassed

Apple's FaceID by putting tape and glasses on unconscious people and transferring their funds through a mobile payment application at Black Hat USA 2019 [164]. They claimed a weak point of the FaceID algorithm is that it did not extract 3D information from the eye area when recognizing the glasses.

Therefore, Presentation Attack Detection (PAD) technique (also called anti-spoofing detection), i.e. automated determination of a presentation attack, plays an important role in securing recognition systems from PAs. At the early stage, most existing PAD algorithms were designed based on the traditional handcrafted features [183] (e.g., Local Binary Pattern (LBP) and Binarized Statistical Image Features (BSIF)) or dynamic signals (e.g., eye-blink and head movement) to capture the attack artefacts. Due to the less representation capabilities of those features, recent works have leveraged deep learning techniques to mine the discriminative clues between bona fide and attack samples and showed great progress in PAD performance [93, 168, 250, 251]. Recent iris [61, 243] and face [13] PAD competitions demonstrated that CNN-based PAD algorithms achieved impressive performance under known/seen scenarios. However, the reported results also indicated a significant PAD performance degradation to unknown variations, such as unknown attacks, sensors, and environments.

This chapter will first present the motivation for generalized PAD solutions in Section 1.1. Then, this is followed by stating the research questions on which this dissertation is based in Section 1.2. Finally, Section 1.3 presents an outlook for the next chapters.

1.1. Towards presentation attack detection

PAD is an unsolved research problem for the following reasons. First, PAD is a self-evolving problem, where PAs and PAD algorithms evolve iteratively, making it very challenging. For example, due to the COVID-19 pandemic, the performance of PAD algorithms designed for non-occluded faces significantly dropped when dealing with masked faces [84]. This calls for the improvement or development of new PAD algorithms to overcome the masked faces problem. Second, features that discriminate between bona fide and attack samples are mostly intrinsic and subtle to varying degrees. Third, PAD is an open-set problem in the real-world scenario. However, most existing PAD algorithms worked on developing and evaluating under controlled settings (i.e., known data) and did not generalize well on novel PAs and unknown domains. Very limited existing PAD works [159, 200, 203, 204] addressed this uncertain issue between development (training) and deployment (testing) scenarios and proposed several cross-domain PAD solutions. Moreover, fairness, as part of generalizability in PAD, is extremely understudied.Unlike the conventional generalizability on unknown attack types or capture environments, the fairness-related generalizability in PAD is concerned with uncertain identity properties, covering unknown demographic and non-demographic attributes. As biometric recognition is gradually reaching into society and our daily life, this raises concerns about the ethical issues of adopting such biometric systems. Fairness is a recently established area of machine learning and refers to the process of correcting and eliminating algorithmic bias from machine learning models. Machine-learning-powered biometrics can lead to unfair treatment of individuals in certain demographic groups based on their gender, age, and race. For example, Face Recognition (FR) algorithms are known to be unfair between certain demographic and non-demographic groups [8, 64, 69, 208, 216, 218]. However, fairness in PAD is unclear.

To address the above issues, this thesis focuses on the generalizability of PAD on unknown variations. In recent years, iris and face recognition are gaining in popularity. Due to the high reliability of iris recognition, it can be used in scenarios where a high level of security is required, such as in banking and financial or government organizations, to deny access to unwarranted identities. Face recognition has developed rapidly and has been widely deployed in our daily life, especially with the advancement of mobile devices. Considering their increased popularity, high accuracy, and high security demand, this thesis is concerned with the generalizability in iris and face PAD. In the case of fairness-related generalizability, this thesis only addresses fairness in face PAD and neglects it in iris PAD. The reasons for this are: 1) A face carries more properties, such as face shape, hairstyle, and accessories, in comparison with iris. 2) Attributes of iris samples are very hard to label by human experts. 3) Compared to face PAD datasets, iris PAD datasets are relatively small-scale and lack data diversity.

To bridge such under-explored gaps, the research efforts in this thesis are categorized into three aspects: fairness assessment and enhancement in face PADs, the emerging challenge of masked face PAD during the COVID-19 pandemic, and generalizability enhancement of iris and face PAD systems. These three aspects will be studied by tackling a number of research questions raised by the developed goals of generalized PAD and offers solutions that address these research questions.

1.2. Research questions

To bridge the above-discussed gaps in current State of The Art (SOTA), this thesis aims at boosting the fairness and generalizability of PAD through addressing a set of unsolved research questions. In order to put these questions in a broader perspective and to provide topic-specific answers, they are divided into three groups based on the research areas. The first group focuses on fairness assessment and enhancement in face PAD. The second group deals with the emerging challenge of masked face PAD. The third is concerned with the generalizability of face and iris PAD. The rest of this section presents the sets of detailed research questions following the three groups, in Sections 1.2.1, 1.2.2 and 1.2.3, respectively.

1.2.1. Fairness assessment and enhancement

Prior studies [64, 99, 115, 202] pointed out that biometric recognition algorithms exhibit discriminatory behaviours against certain demographic groups. The 2019 National Institute of Standards and Technology (NIST) Face Recognition Vendor Test [99] also demonstrated that all 106 tested face recognition algorithms exhibit varying unfair performances on gender, race, and age groups of a mugshot dataset. Creager *et al.* [47] stated that deploying biased recognition systems to law enforcement is potentially unethical. To avoid the potential negative societal impact, assessing the fairness of biometric systems and developing a fair biometric system is an important research direction. However, unlike growing numbers of fairness studies in face recognition, the fairness of PAD, and thus of a major aspect in the biometric system security, is overlooked.

Therefore, in the first research area, this thesis focuses on the fairness assessment and enhancement in face PAD. The questions listed in the following aim to address the clear gap in analysing the fairness in PAD, as well as in providing solutions to enhance this fairness.

- **Fairness assessment:** Building a PAD system involves development data collection and pre-processing, model development, model validation and testing. Each component may exhibit potential unfairness as defined in [99]. Prior fairness research in recognition algorithms [227] indicated that the imbalance of the development data is one of the triggers of unfairness, where models tend to learn better representations for the majority group in the dataset. The decision threshold applied on the output of the PAD models are chosen in the validation stage to fit deployment requirements, this choice might additionally cause further unfairness. This leads to the first research questions tackled in this thesis, and it is stated as follows:

  RQ1: Are PAD systems fair? And does the lack of balance in PAD training data and the methodology of choosing PAD decision threshold affect this fairness?

- **Fairness enhancement:** This thesis uncovers different aspects of unfairness in face PAD. Such unfairness in recognition algorithms motivated a diverse set of fairness enhancement solutions [25, 95, 209, 219], however, no such solutions previously addressed fairness of PAD algorithms. Knowing that motivated this work to aim at innovating a technical solution to enhance the fairness of face PAD by addressing the following question:

RQ2: Is the fairness of PAD solutions enhanced when controllably augmenting the training data so that different data groups will posses specific properties of other groups?

In an effort to answer these questions, this thesis first provides manually annotated attribute labels for a combined face PAD dataset. Then, the fairness of PAD systems is assessed based on four PAD solutions, including traditional handcrafted features and deep-learning-based methods. Finally, this thesis presents a cross-identity and cross-attribute patch swap technique to enhance the fairness of face PAD algorithms. This approach can serve as a plug-and-play data augmentation technique to disorder the identity or semantic information and thus guide the model to mine attack clues, resulting in a fairer face PAD system.

1.2.2. The emerging challenge of masked face PAD

Due to the SARS-CoV-2 coronavirus outbreak, wearing a face mask has become one of the most efficient ways to protect and prevent getting infection. Several studies explored the effect of face masks on the performance of face recognition verification [51, 54, 181] and their results have shown that face recognition algorithms suffer performance degradation because of the face masks. However, the effect of masks on the PAD behaviour and the FR vulnerability to PAs have so far been overlooked. Driven by the ongoing pandemic and different hygiene requirements, PAD systems have encountered several understudied challenges when facing masked faces: 1) unclear vulnerability of FR systems to masked PAs, 2) uncertain relationship between the performance of PAD techniques and PAs with face masks, 3) no solutions to target the masked face PAD problem.

Therefore, the second research area covers these emerging challenges, and the following research questions are drawn to tackle the above-mentioned masked face attack problems.

- **Vulnerability of recognition models:** Several previous studies evaluating the effect of face masks on face recognition performance reported that genuine score distribution (i.e., distribution of scores obtained by comparing bona fide references and probes belonging to the same identity) was significantly affected, and zero-effort-imposter score distribution (i.e., distribution of scores obtained by comparing bona fide references and probes belonging to the different identities) did not seem to be strongly affected by masked probes [51, 54, 181]. The genuine score distribution strongly shifted towards the imposter score distributions. However, the effect of face masks on face recognition vulnerability to presentation attacks has not been investigated, i.e., the effect on the PA comparison score to the bona fide reference when either the bona fide or PA is masked. PA comparison score distributions is obtained by comparing bona fide references and attack probes belonging to the same identities. This raises the following question:

RQ3: Is the vulnerability of FR systems to PAs affected by wearing a mask?

- **Behaviour of PAD models:** The detection of various PAs has been comprehensively addressed in existing PAD research, ranging from 2D attacks such as print and replay attacks to 3D mask attacks. As a result of the COVID-19 pandemic, PADs are being challenged by processing unfamiliar face presentations, i.e. masked faces. This also includes the possibility of a masked PA. Masked face PA in this thesis refers to the print or replay presentation attack, in which the subject is wearing the face mask, or a real face mask is placed on such attacks to simulate wearing a face mask The effect of a face mask on face PAD behaviour has not been yet studied and thus leads to the next question:

  RQ4: Is the behaviour of existing PADs effected by wearing face masks, whether on bona fide or attack faces?

- **Enhancing the PAD of masked faces:** After analysing the effect of wearing a mask on PAD behaviour, a solution is need to minimize this unwanted effect. To mine the subtle and fine-grained local features between bona fide and attack samples, recent PAD studies have tended to utilize pixel-wise supervision strategies during the training phase and have significantly improved performance. However, the values on pixel-wise ground-truth for attack samples are sub-optimally assumed to be the same, either on a 2D binary or 3D depth ground-truth and especially when facing structured samples such as masked faces. This leads to the next research question in this thesis:

  RQ5: Can the PAD performance, especially on masked samples, be enhanced by designing a structured PAD training supervision strategy that takes the possibility of a masked face into account?

To answer the above questions, this thesis first presents a novel Collaborative Real Mask Attack Database (CRMA) dataset, including three types of PAs and both, masked and unmasked, bona fide samples. Then, an in-depth vulnerability analysis of FR systems to such attacks is presented by evaluating three deep-learning-based FR techniques on three types of PAs. Furthermore, extensive experiments are conducted to explore the effect of masked faces on the face PAD behaviour. Seven face PAD algorithms are selected to evaluate the performance and generalizability in intra- and cross-database scenarios under three mask-related protocols. Finally, a PAD solution is successfully proposed to target the masked face PAD problem by refining the training supervision strategy and the inference process. The proposed solution is not directly linked to specific network architecture and thus can be directly incorporated into any common or custom-designed PAD frameworks.

1.2.3. Generalizability of PAD

In recent years, many works have leveraged deep learning techniques and showed great progress in iris [37, 89, 155, 205, 238, 242] and face [93, 159, 168, 200, 200, 203, 204, 250, 251, 251] PAD performance. Despite the huge success of deep learning PAD models under the general seen scenario, these models still show a huge performance degradation when facing unknown variations such as novel types of attacks, attacks in unknown environments, and attacks captured by unknown sensors. Most existing iris and face PAD works are competing to boost the performance under the intra-dataset scenarios, while the generalizability of the PAD systems on unknown scenarios, representing the real use-case, is still an open issue.

The following research questions are stated to aim at enhancing the generalizability of PAD systems. The questions listed in the following aim to enhance the generalizability of face and iris PAD, given that they are some of the most widely used biometric characteristics [48, 122, 198, 248].

- **Face PAD:** The recent face PAD competition [13] evaluated PAD algorithms on unseen face attacks and reported a significant performance degradation under unknown scenarios. This suggests that continuous efforts are needed to improve PAD algorithms to detect fast-evolving PAs. Face PAD algorithms can be categorized into texture-feature-based, deep-learning-based, and hybrid solutions. Texture features, such as features extracted from the frequency domain [35, 141], are less affected by the illumination variations, while the deep features are more discriminative between bona fide and attack samples. However, most texture features derived from the frequency domain are sub-optimal because frequency filters are fixed and unlearnable. This motivates the next research question raised in this thesis:

  RQ6: Can the generalizability of face PAD be enhanced by learning to include information from the frequency domain in an optimized manner?

- **Iris PAD:** Similar performance degradation on unknown attacks was observed in a series of iris PAD competitions [61, 243]. One of the main problems is that the features that discriminate between bona fide and PAs are mostly intrinsic, content-irrelevant (e.g., not related to the iris attribute, subject ID), and subtle to varying degrees. Therefore, mining the attack patterns with local-specific and fined-grained details is essential for enhancing iris PAD generalizability. This leads to the next research question:

RQ7: Can analysing spatially aware regions of the iris and its iris/sclera boundary enhance the generalizability of iris PAD?

Following the need to build automatic decisions based on correctly localized clues, attention mechanisms have attracted increasing interest in computer vision tasks [5, 39, 223, 236, 253]. Attention mechanisms endow networks with the ability to focus on more interpretable and robust features by weighing features with different levels of importance to a specific task. This motivates the next research question:

RQ8: Can networks be automatically guided to focus on the attack-discriminant iris region during the PAD training? If so, does this enhance the iris PAD generalization?

Focusing on enhancing the generalizability of face and iris PAD systems, this thesis proposes several PAD solutions to facilitating fine-grained and discriminative feature learning. A learnable frequency composition is presented to complement deep features extracted in the spatial domain and thus improve the generalizability of face PAD systems. To learn subtle iris features, iris regions are spatially segmented and learned to improve the generalizability of iris PAD approaches. Moreover, this thesis takes advantage of the attention mechanism to design a generalized iris PAD network that takes into account the need to focus on attack-discriminant regions.

1.3. This thesis

After stating the motivation and introducing the highlights of this thesis, this section presents an overview of the rest of this work as follows:

Chapter 2 provides the essential background information in terms of the typical structure of a PAD system within a biometric system. First, the main components of biometric recognition and PAD system are presented. Then, evaluation metrics are discussed to measure the performance of PAD systems, the vulnerability of recognition systems, and the fairness of biometric systems. It ends by introducing an overview of deep-learning-based PAD models, including main network components (i.e., layer structure, data augmentation, and attention mechanism) used in this thesis and mainstream iris and face PAD benchmarks.

Chapter 3 targets the under-explored gaps in understanding and enhancing fairness in the practical applications of face PAD. To enable answering *RQ1*, this thesis first combines several face PAD datasets, and seven attribute annotations are labelled, involving demographic and non-demographic attributes. Then, the fairness of PAD systems is assessed from two aspects, the nature of training data and the methodology of choosing PAD decision thresholds, as a response to *RQ1*. To enhance the fairness in face PAD, this thesis further proposes a PAD solution based on carefully designed data augmentation technique to disrupt the demographic/semantic information, in response to *RQ2*. This chapter is based on works [72, 81, 86].

Chapter 4 is concerned with the emerging challenge of the masked face PAD during the COVID-19 pandemic. To enable answering *RQ3* and *RQ4*, a new CRMA dataset is presented in this thesis, including unmasked and masked bona fide and attack faces, and partially masked face attacks (i.e. PA is covered by a real medical facial mask). As a response to *RQ3*, extensive experiments are conducted to explore the effect of masks on the vulnerability of FR by adopting three conventional FR methods. Then, in response to *RQ4*, three protocols are designed to measure the generalizability of the current PAD algorithms on unknown masked bona fide or attack samples by adopting seven PAD algorithms including hand-crafted feature-based and deep-learning-based solutions. Knowing the PAD performance degradation on masked faces, Chapter 4 further presents a structured PAD training supervision strategy to address the masked face PAD, responding an answer to *RQ5*. This chapter is based on the published papers [75, 84].

Chapter 5 focuses on enhancing the generalizability of iris and face PAD systems. First, this chapter proposes to intelligently leverage the information from the frequency domain to mine subtle artefact clues, providing an answer to previously stated *RQ6*. Furthermore, as a response to *RQ7*, spatially aware regions of the iris and its iris/sclera boundary are analysed to enhance the performance and generalizability of iris PAD under unknown scenarios. Lately, aiming to guide model automatically focus on attack-discriminant iris regions, the attention mechanism based iris PAD method is proposed and evaluated under the cross-domain evaluation, including the cross-spectrum scenario, responding to *RQ8*. This chapter is based on the published papers [74, 77, 78, 79, 80, 82, 83]

Chapter 6 concludes this thesis by highlighting its contributions, elaborating its practical benefits, and giving an outlook for future research.

1.4. Summary

This chapter first presented a motivation leading to a set of research questions posed in this thesis. The listed questions are aimed at boosting the fairness and generalizability of PAD. These research questions were based on three targeted challenges, fairness assessment and enhancement in face PAD, the emerging challenge of masked face PAD systems, and the generalizability of PAD systems, as shown in Figure 1.1.

The first focus of the presented research questions was concerned with the fairness in

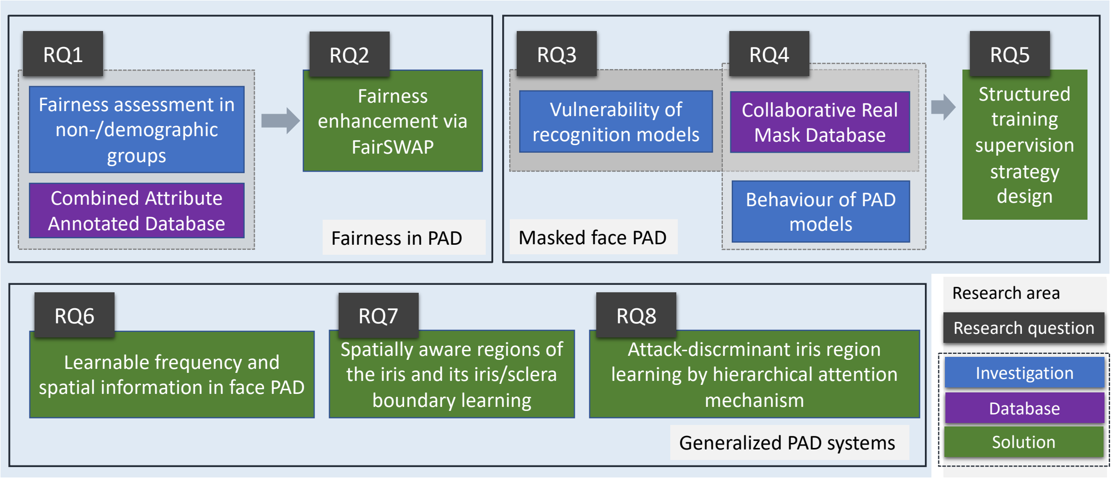

RQ1
Fairness assessment in non-/demographic groups
Combined Attribute Annotated Database
RQ2
Fairness enhancement via FairSWAP
Fairness in PAD
RQ3
Vulnerability of recognition models
RQ4
Collaborative Real Mask Database
Behaviour of PAD models
Masked face PAD
RQ5
Structured training supervision strategy design
RQ6
Learnable frequency and spatial information in face PAD
RQ7
Spatially aware regions of the iris and its iris/sclera boundary learning
RQ8
Attack-discrminant iris region learning by hierarchical attention mechanism
Generalized PAD systems
Research area
Research question
Investigation
Database
Solution

Figure 1.1.: An overview of the contributions in relation to the research questions posed in this thesis. **Fairness in PAD** focuses on analyzing and enhancing the fairness of face PAD algorithms, **Masked face PAD** targets the generalizability of face PAD to subjects wearing face masks, and **Generalized PAD systems** provides a set of solutions to boost the generalizability of face and iris PAD over the attack and capture variabilities.

PAD systems by offering a comprehensive assessment of fairness in face PAD and proposing an intelligent augmentation technique to enhance fairness. The second group of research questions focused on addressing the emerging challenges of the masked face PAD by providing an in-depth analysis of the impact of masked faces on PAD performances and FR vulnerabilities. As a result, a structured training supervision strategy is designed to address the observed performance degradation on masked faces. The third group of research questions targeted the generalized iris and face PAD solutions. Three PAD methods were proposed to boost the generalizability of PAD systems under unknown scenarios, such as unknown attacks, unknown capture sensors and environments.

2. Background

The previous chapter presented motivation and a structure for the research problems addressed in this thesis. This chapter presents crucial background information and definitions to facilitate a better understanding of the following chapters. First, the standard definitions for biometric systems and PAD systems, along with their main components, are introduced. Then, this chapter presents the biometric performance metrics that are commonly used in the literature, including the PAD performance and fairness metric. Several deep-learning-based techniques, including main layers of Convolutional Neural Network (CNN), data augmentation techniques, and attention mechanisms are introduced. Finally, this chapter includes more detailed insights into deep-learning-based PAD networks and the mainstream datasets.

2.1. Biometric systems with PAD

Biometric characteristic, e.g., fingerprint, iris, and face [131], is defined as a biological and behavioural characteristic of an individual. A biometric characteristic is a distinct and repeatable biometric feature that can be extracted for the purpose of biometric recognition. In the last decade, biometric systems have been incorporated with authentication tools and are applied in many scenarios, such as identity access management and access control, surveillance, and security [1, 133]. Moreover, the rapid evolution of biometric recognition systems has raised concerns about their ability to resist PAs. This section will introduce the components of biometric recognition systems with PAD systems.

A biometric recognition system is a technology that takes an individual biometric trait as input, analyzes, and verifies or identifies the individual, while PAD is a technique for the automated detection of PAs. This subsection will first present a generic biometric recognition system based on the definition in ISO/IEC 19795-1 [132]. Then, a PAD system which can integrated into recognition system is introduced following ISO/IEC 30107 [126] standards and illustrated in Figure 2.1.

A generic biometric recognition system can be briefly divided into five main subsystems: data capture, signal processing, data store, comparison, and decision subsystems, as shown

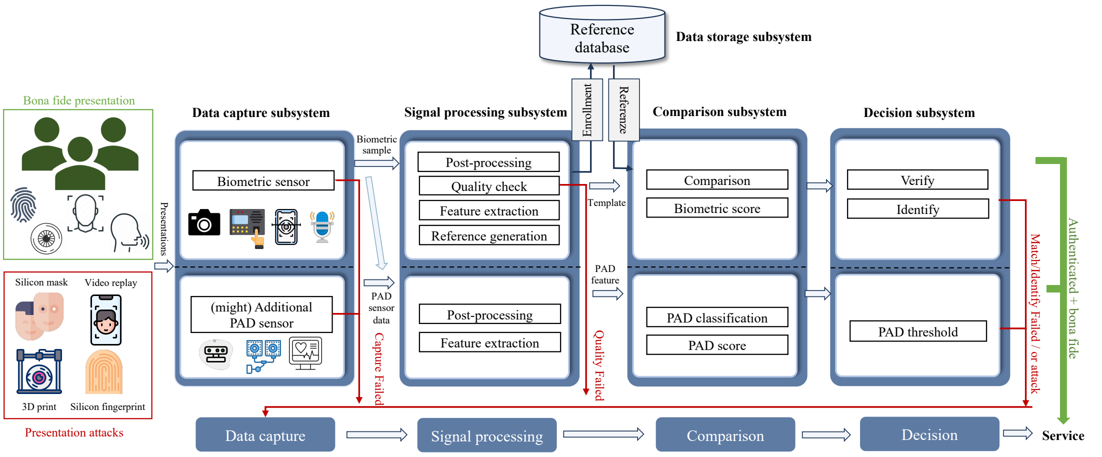


Figure 2.1.: The pipeline of biometric systems including PAD.

in Figure 2.1. First, the biometric capture devices (e.g., camera, fingerprint scanner, or iris scanner) captures a signal from the presented biometric modality. Once the biometric samples are successfully captured, they will be sent to the signal processing subsystem. The signal processing subsystem can couple quality control module that assesses the suitability of the biometric sample for recognition [131, 217]. If the quality of the captured biometric sample is insufficient for further processing steps, the biometric characteristic shall be presented again. In general, a generic recognition signal processing subsystem consists of the post-processing of biometric samples (e.g., enhancement and segmentation), quality check, feature extraction, and reference generation. The output of the biometric signal processing subsystem is the identity template which is a distinctive and repeatable feature. For the enrolment step, the generated identity template is then transferred to the data storage subsystem and stored in the enrolment dataset in accordance with ISO/IEC 24745 [2]. For further verification and identification steps, a comparison subsystem compares the biometric probe produced by the signal processing subsystem of biometric reference(s) enrolled by the genuine data subject(s). In the case of verification, a single biometric probe is compared to a signal reference. When performing identification, a probe is compared to all or a subset of references. The comparison subsystem produces the comparison score(s) and transmits it to the decision subsystem. The comparison score refers to the similarity score in the recognition system. If the received comparison score is higher than a pre-defined threshold (and/or the comparison score is ranked within a pre-defined number of ranked values), the verification (identification) is successfully performed.

However, it has potential security problems when passing such authenticated users without attack detection to the various services. A potential attacker can use various attack operations to target the biometric recognition system, such as the presentation of artefacts to the capture device, direct manipulation of the underlying algorithms like template creation, and modification of biometric templates in the enrolment dataset. In this thesis, we focus on presentation attacks. As defined in ISO/IEC 30107-3 [127], a presentation attack is the attempt to present a biometric characteristic to the biometric data capture subsystem with the goal of interfering with the operation of the biometric system. Considering that data from the biometric sensor (e.g., face images) may be insufficient to conduct a PAD task, biometric systems with PAD mechanisms might contain additional sensors to detect specific properties of a biometric characteristic, such as depth sensor and PhotoPlethysmoGraphy (PPG) detector. A biometric recognition system can couple PAD systems as shown in Figure 2.1. After successfully capturing the biometric sample, this is transmitted to the PAD signal processing subsystem, which involves the post-processing of biometric samples and PAD feature extraction modules. Unlike in recognition systems where recognition performance is usually linked directly to biometric data quality, quality control might be deactivated by an attacker. For example, samples from attack artefacts can exhibit better quality than samples from bona fide biometric characteristics. Therefore, the PAD signal processing subsystem contains no quality check in most cases. The output of the PAD signal processing subsystem is the extracted PAD feature which is then sent to a PAD comparison subsystem to perform the classification and produce a PAD score. Furthermore, the produced PAD score is compared with a pre-defined PAD threshold. If the biometric presentation is detected as an attack, this biometric sample may be presented again. Lastly, a negative PAD result (no PA detected) is required for a final successful authentication. After passing the biometric recognition and PAD systems, the user can access services, such as unlocking devices, financial access, and automated gate access.

2.2. Biometric performance metrics

This section introduces performance metrics for evaluating the PAD system and assessing the vulnerability of the biometric recognition system following the definition in ISO/IEC 30107-3 [127]. As some PAD works presented in the literature did not follow the ISO/IEC 30107-3 [127] terms for PAD performance evaluation, this section also provides the main evaluation metrics in the literature for the sake of comparability. Lastly, this section presents the metrics for assessing the fairness of the decision-making based biometric systems. There is no international standard metric definition and no study for the fairness assessment of PAD systems yet. Therefore, the introduced fairness metrics are based on reported metrics in the recognition literature and adapted to the PAD systems.

2.2.1. PAD metrics

The ISO/IEC 30107-3 [127] established the principles and metrics for performance assessment of glsPAD mechanisms. In ISO/IEC 30107-3 [127], evaluation of PAD mechanisms is categorized as three levels: PAD subsystem, data capture subsystem, and full system. A PAD subsystem refers to develop a PAD mechanism and make an explicit declaration regarding the detection of PAs. A data capture subsystem including capture hardware and/or software couples the glsPAD mechanism and quality checks. Capture subsystem may not know whether the presented data is bona fide or attack. A full system adds biometric comparison to the PAD subsystem or data capture system, comprising a full end-to-end system (as shown in Figure 2.1). This leads to additional failure points for the Presentation Attack Instrument (PAI) beyond PAD mechanisms and quality checks. In a full system, there might be one or multiple glsPAD mechanisms at different points in the system. In most of existing PAD literature, evaluation of PAD mechanism and resulting reports are based on the PAD subsystem level. The main used evaluation metrics in this level are Attack Presentation Classification Error Rate (APCER) and Bona fide Presentation Classification Error Rate (BPCER). APCER is the proportion of attack presentations using the same PAI species incorrectly classified as bona fide presentations in a specific scenario. BPCER is the proportion of bona fide presentations misclassified as presentation attacks in a specific scenario. Both APCER and BPCER metrics are computed based on the system operation threshold calculated from the development data that control the trade-off between these metrics. Additionally, Average Classification Error Rate (ACER) or Half Total of Error Rate (HTER) in some literature [14, 155], the average value of APCER and BPCER, is used to report the overall PAD performance. In addition to standard metrics in ISO/IEC 30107-3 [127], it is common to report the PAD performance in terms of single figure BPCER at fixed APCER or APCER at fixed BPCER for demonstrate the PAD performance in practice [205]. For instance, Sharma *et al.*[205] reported True Detection Rate (TDR) (equals to 1 - APCER) at a false detection rate (equals to BPCER) of 0.2%. Another two common metrics to report the overall performance for comparability are Equal Error Rate (EER) and Correct Classification Rate (CCR). EER is the APCER or the BPCER at the operation point where they are equal, and CCR is the proportion of correctly classified samples among all samples. To provide a visual evaluation, Detection Error Tradeoff (DET) or Receiver Operating Characteristic (ROC) curves are used to report PAD performance at all operational points, showing security measures versus convenience measures. DET curve plots BPCER (y-axis) vs. APCER (x-axis), while ROC curve plots

(1-BPCER) (y-axis) vs. APCER (x-axis). Several PAD literature also opt to include the area under ROC curves as evaluation metric [139, 203]. In addition to the PAD subsystem evaluation level, the most used metric in a full system evaluation of a verification system is Imposter Attack Presentation Match Rate (IAPMR), which is the proportion of impostor attack presentations using the same PAI species that result in accept, as defined in ISO/IEC 30107-3 [127].

2.2.2. Biometric fairness metrics

In recent years, deep learning techniques have achieved great success in various biometric tasks, including face/iris/fingerprint recognition and presentation attack detection. Despite the improved performance, decision-making based has been shown to exhibit discriminatory behaviour against certain groups [4, 29, 46, 81, 218]. As "fairness" and "bias", among other terms, are being used to describe performance variations across data groups in biometric studies and generally in AI, our stand on the terminology and choice to use the term "fairness" in this thesis is shortly explained in the following. The systematic difference in treatment of certain objects, people, or groups in comparison to others in AI is referred to as Bias in AI according to recent AI standardization document (ISO/IEC TR 24027:2021 [128]) released by the ISO/IEC JTC 1 Information technology Subcommittee SC 42 on Artificial intelligence. The same document would use the term "Fairness" and "Bias" as representation of AI bias and discuss the possibility of assessing fairness in AI with a "fairness measure". Biometric literature adapts the terms of Bias and Fairness and their measures from AI standards and the broader machine learning and artificial intelligence literature. Recent efforts to standardize quantifying biometric performance variation across demographic groups within the ISO/IEC JTC 1 Information technology Subcommittee SC 37 on Biometrics are underway. After discussions with some of the developers of this standard, which is still under development (ISO/IEC WD 19795-10 [129]), the adaption of fairness measures from AI is seen as lacking the consideration of social, legal, or cognitive aspects of fairness, which we agree with. Therefore, the standard is being developed in the direction of avoiding the confusion with such aspects, and thus using the term "aggregate equitability measure" in lieu of the term fairness measure [129]. Our work supports the use of this term and the motivation behind it. However, given that this standard is still under development (not released) and may still be edited, we opted, for now, to follow the common terminology in biometric literature and AI standardization (ISO/IEC TR 24027:2021 [128]) and discuss the biometric performance variation across different groups under the "fairness" term. A fairness metric, named Fairness Discrepancy Rate (FDR), was proposed [64]. FDR is proposed to evaluate and compare the fairness aspects between different biometric verification systems. FDR metric exhibits a better representation by leveraging a single ’fair’ decision threshold from all test groups than other fairness metrics, such as using DET curve or ROC curve, because DET curve or ROC curve comparison was conducted assuming that decision thresholds are group-specific, which is not feasible in a real-world scenario. A PAD system is considered fair if different demographic or non-demographic groups share the same APCER and BPCER for a given decision threshold τ, where $\tau = APCER_x$ from all groups. The FDR in [64] is adapted to PAD performance as follows:

$$A(\tau) = max(|APCER^{d_i}(\tau) - APCER^{d_j}(\tau)|), \quad \forall d_i, d_j \in D \tag{2.1}$$

$$B(\tau) = max(|BPCER^{d_i}(\tau) - BPCER^{d_j}(\tau)|), \quad \forall d_i, d_j \in D \tag{2.2}$$

$$FDR(\tau) = 1 - (\alpha A(\tau) + (1 - \alpha)B(\tau)) \tag{2.3}$$

where D is a set of attribute groups $D = \{d_1, d_2, ..., d_n\}$, α is a hyper-parameter to weigh those differences, representing the level of concern applied to differences between APCER (referring to False Match Rate (FMR) in [64]) and BPCER (referring to False Non-match Rate (FNMR) in [64]). The value of FDR varies from 0 (maximum discrepancy) to 1 (minimum discrepancy). As a result, FDR has smaller-is-unfairer semantics. In addition to FDR, Chapter 3 will present an accuracy balanced fairness that concerning both the PAD performance and fairness of PAD systems.

CNN plays a vital role in the deep learning technique, especially in varied computer vision tasks, including biometric traits PAD tasks. Therefore, the following section introduces the details of three main types of layers used in CNN architecture.

2.3. Deep learning techniques

2.3.1. Convolutional neural networks

Convolutional neural networks are a class of deep neural networks (DNN). CNN has been tremendously applied to various computer vision tasks such as FR, image classification, object detection, and image segmentation. CNN architectures have yielded excellent success in various computer vision tasks, such as image classification, object detection, face/iris/fingerprint recognition and PAD. CNN take advantage of the fact that inputs are images in most cases and thus constrain the network architecture in a more sensible way. A typical CNN architecture is formed by stacking three main types of layers: Convolutional layer, Pooling layer, and Fully-connected layer. The following is an overview of the three main components.

Convolutional Layer: The convolutional layer is the core building block of CNN and does most of the computational heavy lifting, which computes the convolutional operation of the inputs to extract fundamental features. The convolutional layer consists of a set of learnable filters (or kernels) with small spatial width and height size and large depth extended from input. During the forward pass, each filter is convolved with the input images/feature map, and an activation map is created. Specifically, each filter is slid over the width and height of the input volume, and a dot product is computed between the entries of the filter and the input at any position. The convolve process and size of the output volume are controlled by three hyper-parameters: depth, stride, and padding. The depth of output volume corresponds to the number of used filters. Stride hyper-parameter refers to the step size with which the filter slides. When the stride is one, the filter is moved one pixel at a time. Padding allows us to control the spatial size of the output volume by filling with a specific number, e.g., zeros around the border. Overall, the convolutional layer learns features by having relatively few parameters due to the sparse connections and parameter sharing. On the one hand, sparse connections refer to that each filter is connected to only a local region of the input volume (named receptive field). On the other hand, parameter sharing could significantly reduce the number of parameters by using a single filter to slide across all receptive fields. Therefore, CNN is able to capture the spatial relations between separate features during convolutional operation.

Pooling layer: To reduce the amount of parameters and computation in the CNN, pooling layer is used to reduce the spatial size of the input feature map progressively. Consequently, it is common to insert a pooling layer in-between successive convolutional layers. The pooling layer operates independently on each channel (depth) of the input and resizes it spatially. i.e., depth dimension remains unchanged. Two types of pooling in popular use are max (Max-Pooling) and average pooling (Avg-Pooling). Max-pooling uses the maximum value of each local region in the feature map, while Avg-Pooling computes the average value. Similar to the convolutional layer, the pooling layer requires two hyper-parameters: pooling size and stride. It should be noted that the pooling layer is not trained during the backpropagation of gradients, as the output volume only depends on the values of the input volume and hyper-parameters.

Fully Connected Layer: The fully connected layer connects each neuron to all activations in the previous layer. The activations can thus be computed with a matrix multiplication followed by a bias offset. In the classification task, including the PAD task, several fully connected layers are stacked to extract features and the last fully-connected layer is usually used for classification.

Followed by the presented main layers of CNN, the next section describes several deep-learning-based PAD algorithms and datasets that are popular in use.

2.3.2. Data augmentation

Despite the remarkable performance of deep learning models achieved on many computer tasks, these models are heavily reliant on large-scale and diverse data to avoid overfitting. However, collecting large-scale data is particularly tedious and labor-intensive, especially when faced with biometric data, which is associated with privacy concerns. Therefore, data augmentation, a data-space solution to the problem of insufficient data attracted much attention. Data augmentation technique can increase the diversity of training data by geometric transformation (e.g., flipping, rotation, or translation) [215], color space transformations [215], kernel filters (e.g., Gaussian blur), mixing images [124, 254] and GAN-based augmentation [96]. Moreover, data augmentation technique helps to alleviate the bias problem induced by the training data. This thesis proposed a data augmentation specially designed for PAD task that creates more complex training data by swapping regions between different samples. To some extent, data augmentation enhanced the PAD performance, but PAD algorithms are more sensitive to the selection of the augmentation strategies as explored in [80] than general image classification tasks.

2.3.3. Attention mechanism

Attention mechanism in neural networks mimics the human cognition process of selectively focusing on a few relevant things while neglecting others. The attention mechanism emerged first as an improvement over the encoder-decoder based neural machine translation in natural language processing [5]. Later on, this mechanism and its variants were used in other fields, including computer vision. The significance of attention mechanism has been studied extensively in the previous literature [39, 223, 236, 253]. One of the widely used attention mechanisms is Convolutional Block Attention Module (CBAM) [236] that contains two sub-modules: channel attention module and spatial attention module. The authors [236] pointed out that CBAM can be integrated at any convolutional block to obtain subsequent refined feature maps $\mathbf{F}_{out} \in \mathbb{R}^{C\times H\times W}$ from input intermediate feature maps $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$. First, a channel attention map $\mathbf{M}_c \in \mathbb{R}^{C\times 1\times 1}$ is produced by exploiting the inter-channel relationship of features by following the equation:

$$\mathbf{M}_c(\mathbf{F}) = \sigma(\mathbf{W_1}(\mathbf{W_0}(\mathbf{F}^c_{avg})) + \mathbf{W_1}(\mathbf{W_0}(\mathbf{F}^c_{max}))), \tag{2.4}$$

where σ refers to the sigmoid function, $\mathbf{W_0} \in \mathbb{R}^{C/r\times C}$ and $\mathbf{W_1} \in \mathbb{R}^{C\times C/r}$ denotes the mutil-layer perceptron weights with reduction ratio r and are shared for both inputs. $\mathbf{F}^c_{avg}$

and $\mathbf{F}^c_{max}$ denote average-pooled features and max-pooled features, respectively. Then, a spatial attention map $\mathbf{M}_s \in \mathbb{R}^{H \times W}$ is produced by utilizing the inter-spatial relationship of features. This attention map encodes where to emphasize or suppress by a convolutional operation and and can be formulated as following:

$$\mathbf{M}_s(\mathbf{F}) = \sigma(f^{n \times n}([\mathbf{F}^s_{avg}; \mathbf{F}^s_{max}])), \tag{2.5}$$

where σ denotes the sigmoid function and $f^{n \times n}$ is a convolutional operation with the filter size of $n \times n$. $\mathbf{F}^s_{avg}$ and $\mathbf{F}^s_{max}$ denote average-pooled features and max-pooled features respectively. Finally, two attention modules can be placed in a parallel or sequential manner. Later on, several variant [113, 163, 264] of CBAM were proposed and applied to improve the attention mechanism and thus enhance the performance of models. This thesis also adopted such attention mechanism to enhance the generalizability of the PAD algorithms as described in Chapter 5.

2.4. Presentation attack detection

As biometric systems are widely used in real-world applications, including mobile device authentication and access border control, PAD are becoming a larger threat, where an attacked biometric sample is presented to the biometric system and attempted to be authenticated as introduced in Section 2.1. Therefore, PAD is a very critical step in securing the biometric systems. As shown in Figure 2.1, the captured data samples, such as fingerprint/iris/face images, are sent to the signal processing subsystems. The pre-processing module of the signal processing subsystems consists of a region of interest detection and cropping. Then, the feature extraction module is used to extract distinctive features from cropped samples. Researchers are recently racing to boost performance using deep learning techniques to facilitate discriminative feature learning. Hence, this section focuses on the CNN-based PAD approaches, including backbone architectures, loss function, and the publicly available used face and iris PAD datasets.

2.4.1. Deep-learning-based PAD

CNN have been push the frontier of the computer vision research in the past years. By leveraging deep learning techniques, many works have shown great progress in PAD performance. Most PAD models are built based on the popular network architectures [107, 117, 193, 207] designed for image classification task.

Network architectures

The main network architectures of high-performing face and iris PAD models are designed based on the common CNN architectures used in image classification including VGGNet [207], ResNet [107], DenseNet [117], and U-Net [193]. Some works integrated various modules into the CNN backbones to enhance the feature learning, such as Gated Recurrent Unit (GRU) in [27] and Temporal Shift Module (TSM) in [263]. In addition to extract representative visual features by CNN architecture, some PAD works [168, 245] cooperated the CNN with the Recurrent Neural Network (RNN) architecture, e.g., Long Short-Term Memory (LSTM) [109] to extract temporal correlation across multiple frames. The summarized information of recent face PAD with used backbones, supervision strategy and loss function can be found in the Table 2.1. Moreover, the most used CNN backbones in PAD are briefly described in the following.

VGGNet [207] proposed a very deep architecture by increasing depth with small 3×3 convolution filters. Their experiments showed that a significant improvement on the prior-art configurations can be achieved by pushing the depth to 16-19 weight layers. VGGNet secured the first and second places in the localization and classification tracks in ImageNet competition (ILSVRC) 2014, respectively. VGGNet arethe widely adpoted CNN architecture for face PAD [147, 161, 172, 263] and iris PAD [77, 182, 238].

ResNet [107] is a residual framework to ease the training of networks that are substantially deeper than those used previously. The network architecture is based on an identity shortcut connection (residual connection) that skips one or more layers. The skip connection allows the information to flow by adding the input of a residual block to its output and passing it to the following residual block. With the help of the skip connections, ResNet won first place on classification and localization tasks in the ILSVRC 2015 competition, as well as detection and segmentation in the COCO 2015 competitions. Consequently, ResNet is the most widely used backbones in the face PAD [27, 33, 38, 83, 88, 162, 245, 247, 259] and iris PAD [32].

DenseNet [117] proposed to connect each layer to every other layer in a feed-forward manner to alleviate the vanishing-gradient problem and strengthen feature propagation. For each layer, the feature maps of all preceding layers are used as inputs and its own feature maps are used as inputs into all subsequent layers. DenseNet achieved high performance on four highly competitive object recognition benchmark tasks (CIFAR-10, CIFAR-100, SVHN, and ImageNet) with less computation. Therefore, DenseNet is one of the widely used backbones in the face PAD [93, 180] and iris PAD [37, 79, 205, 242].

U-Net [193] was developed first for biomedical image segmentation. This architecture consists of a contracting path to capture context (also called encoder) and a symmetric expanding path to enable precise localization (i.e., decoder). U-Net won the ISBI cell tracking challenge 2015 in the most challenging categories (Phase contrast and DIC microscopy) by a large margin. Up to now, it is one of the most popularly used approaches in any semantic segmentation task. U-Net architecture is applied in face PAD algorithms as generative model backbone to learn attack artefacts in collaboration with pixel-wise supervision, such as [42, 88, 170, 230] in face PAD.

Loss functions

In the context of the deep learning neural network, the loss function, which is used to evaluate a candidate solution (i.e., weights of the network) plays an important role in guiding the training direction. In most cases, PAD is treated as a classification task (i.e., attack or not). As shown in Table 2.1, it can be noted that most PAD approaches are trained with Cross-Entropy (CE) including Binary Cross-Entropy (BCE) when the model training is supervised only by binary label. CE loss increases as the predicted probability diverges from the ground truth label. To capture the more discriminative artefacts feature, several face PAD approaches [27, 147, 168, 246, 249, 251, 263] trained model with the help of other supervision strategy, such as binary mask, depth map and reflection map. In particular, each pixel in a binary mask is considered a label for small patches in an image. Depth map faces are generated only for bona fide samples, as planar attacks (e.g., print/replay attacks) are considered to contain no depth information. In contrast to the depth map, the reflection map is generated only for the attack sample due to the observable surface reflection of the replayed attack. Consequently, the regression loss is employed, such as L1 loss and Mean Squard Error (MSE) loss, to incorporate with such supervision strategies. L1 loss measures the mean absolute error between each element in the prediction map and ground truth, while MSE measures the average of squared difference between predictions and targets. The generative based PAD model [42, 88, 170, 230] is commonly trained by pixel-wise supervision (e.g., input RGB image or spoof map) and thus most training is based on minimizing reconstruction loss (i.e., regression loss) and adversarial loss.

Method	Year	Backbone	Supervision	Loss
SpoofNet [178]	2015	Custom CNN	Binary	BCE
FASNet [172]	2017	VGG16	Binary	BCE
Auxiliary [168]	2018	custom CNN+LSTM	Depth, rPPG, spectrum	MSE
BASN [147]	2019	VGG16	Depth, reflection	MSE, CE
STASN [245]	2019	ResNet50+LSTM	Binary	BCE
TSCNN [38]	2019	ResNet18	Binary	BCE
Reconstruction [42]	2019	U-Net	Binary mask, RGB input	Structural Similarity (SSIM), regression loss
PixBis [93]	2019	DenseNet161	Pixel-wise binary mask	BCE
CompactNet [161]	2020	VGG19	Binary	Triplet loss
CDCN [251]	2020	Custom CNN	Depth	MSE, contrastive depth loss
DRL-FAS [27]	2020	ResNet18,GRU	Binary	BCE
BCN [246]	2020	Custom CNN	Binary mask, depth, reflection	MSE, CE
Disentangled [257]	2020	Custom CNN	Texture map, depth	Regression loss, adversarial loss, CE
AENet [259]	2020	ResNet18	Depth, reflection	CE
LGSC [88]	2020	ResNet18, U-Net	Binary, binary mask	Triplet loss, regression loss, BCE
STDN [170]	2020	U-Net, PatchGAN	Binary mask, RGB input	adversarial loss, regularizer loss, regression loss
PS [247]	2020	ResNet50/CDCN	Binary mask/depth	CE/contrastive depth loss
TAE [180]	2020	Autoencoder, DenseNet161	Binary, RGB input	MSE, BCE
CIFL [33]	2021	ResNet18	Binary	Focal loss
PCGN [162]	2021	ResNet101	Binary	CE
DAM [263]	2021	VGG16,TSM	Depth	MSE, symmetry loss
DC-CDN [249]	2021	CDCN	Depth	Contrastive depth loss, MSE
Dual-stage disentanglement [230]	2022	ResNet18, U-Net	Binary mask, RGB input	Adversarial loss, MSE, triplet loss, CE
LMFD-PAD [83]	2022	ResNet50	Binary mask	Focal loss, BCE

Table 2.1.: Overview of the recent face PAD algorithms. Recent commonly SOTA face PAD algorithms utilized the ResNet [107] as network backbone and training was supervised by the binary label. Several approaches also tend to use other supervision strategies, such as binary mask, depth map, and reflection map to boost the PAD performance. As a result, the selection of loss is dependent on the used supervision strategy.

2.4.2. Evaluation datasets

This section presents the mainstream face and iris PAD used in the literature as well as in this thesis.

Face PAD datasets

Table 2.2 summarizes the main publicly available face PAD datasets used in the literature. The datasets that are used in this thesis to evaluate the generalizability of face PAD algorithms are also briefly described in the following.

CASIA-FASD [261]: CASIA-FASD [261] dataset covers a diverse potential attack types in real-world including warped photo attack, cut photo attack and video attack. CASIA-FASD includes 600 videos across 50 subjects, where each subject contains 12 videos (three bona fide and nine attacks). The videos consist of low, normal and high quality, where videos of low and normal quality have a resolution of 640×480, and high-quality videos with a resolution of 1280×720.

Replay-Attack [43]: Idiap Replay-Attack [43] dataset contains 1,200 videos (200 bona fide and 1,000 attack samples) from 50 subjects. The bona fide and attack videos were acquired in two sessions: controlled and adverse where the background and illumination condition was not uniform. The attack samples were captured under two modes: hand-based attack (i.e., the attack device was held by a human hand) and fixed-support attack (i.e., the attack device is placed on fixed support).

MSU-MFSD [233]: MSU-MFSD [233] dataset consists of 440 videos (110 bona fide and 330 attack samples) across two attack types, video replay and printed photo, from 55 subjects.

OULU-NPU [14]: OULU-NPU [14] dataset was captured under the realistic mobile authentication scenarios, where a total of 5,940 videos (1,980 bona fides and 3,960 attacks) from 55 subjects was captured under various illumination and background scene by six mobile phones. Four protocols are provided to evaluate the generalizability of algorithms across different covariates. Protocol-1 is designed to study the impact of environmental conditions (e.g., illumination and background scene). Protocol-2 evaluates different attacks created by various instruments. Protocol-3 is a leave one camera out protocol, aiming to examine the effect of different capture mobile phones. Finally, Protocol-4 explores all the challenges above by leave-one-out cross-validation.

HKBU-MARs V2+ [166]: HKBU-MARs V2+ [166] is a 3D mask PAD dataset where the 3D masks were generated by two different techniques. The first technique is the ThatsMyFace mask which leverages 3D reconstruction and 3D printing techniques to generate the facial mask. The advantage of ThatsMyFace Mask is that the mask can be easily generated with a single customized image, but the skin texture is not well restored due to the defects of the 3D printing technique. The second technique is the REAL-F mask which provides a high-quality appearance and looks very similar to a bona fide face as the skin texture, blood vessels of the eyes and iris can be reconstructed. This dataset contains 180 videos (120 bona fides and 60 attacks) from 12 subjects.

WFFD [138]: Unlike most of the above face PAD datasets that focus on replay and print attacks, WFFD [138] dataset provides wax figure faces as super-realistic 3D PAs. It consists of 2,200 images with both bona fides and wax figure attack faces (a totally of 4,400 faces from 745 subjects) with a high diversity from online collections.

CRMA [84]: CRMA database [84] contains: 1) both unmasked and masked bona fide samples [51, 54], 2) conventional replay and print PAs created from faces not wearing a mask, 3) replay and printed PAs created from masked face images, and 4) partial attack where the unmasked printed/replayed faces are covered with real masks. The bona fide data collected by Damer *et al.*[51, 54] were adopted for investigation of the effect of wearing a mask on face verification performance, while the attack samples were provided in this thesis [84] for the exploration of the effect of face mask on face PAD performance. The CRMA database comprises 423 bona fide videos and 12690 attack videos (more details in Chapter 4). CRMA database is a challenging PAD database due to different face masks, multiple capture sensors, and various capture distances.

Iris PAD datasets

Table 2.3 summarizes the main publicly available iris PAD datasets used in the literature as well as used in this thesis for evaluate the iris PAD performance. The detailed information of these datasets is also briefly described in the following.

NDCLD-2013 [143]: The NDCLD-2013 dataset comprises of 5100 NIR images and is conceptually divided into two sets based on capture sensors: 1) LG4000 including 4200

Database	Year	# Sub	# Data (BF/attack)	Capture devices (BF/attack)	Display devices	Modality	Attack type
NUAA [214]	2010	15	5105/7509 (I)	Webcame	-	RGB	1 Print
CASIA-FAS [261]	2012	50	150/450 (V)	Two USB cameras, Sony NEX-5	iPad	RGB	1 Print, 1 Replay
Replay-Attack [43]	2012	50	200/100 (V)	MacBook 13 / iPhone 3GS, Cannon SX150	iPhone 3GS, iPad	RGB	1 Print, 2 Replay
3DMAD [71]	2013	17	170/85 (V)	Microsoft Kinect	-	RGB/Depth	1 3D Mask
Msspoof [45]	2015	21	1,680/3,024 (I)	uEye camera	-	RGB/NIR	1 Print
MSU-MFSD [233]	2015	35	110/330 (V)	MacBook Air, Google Nexus 5 / Cannon 550D, iPhone 5s	iPad Air, iPhone 5s	RGB	1 Print, 2 Replay
HKBU-MARs V2+ [166]	2016	12	120/60 (V)	Logitech C920, industrial camera, EOS M3, 4 mobile phones	-	RGB	2 3D masks
Oulu-NPU [14]	2017	55	1,980/3,960 (V)	6 smartphones	Dell 1905FP, Macbook Retina	RGB	2 Print, 2 Replay
SiW [168]	2018	165	1,320/3,300 (V)	Cannon EOS T6, Logitech C920 webcam	iPad Pro, iPhone 7, Galaxy S8, Asus MB 168B	RGB	2 Print, 4 Replay
CASIA-SURF [258]	2018	1000	18000/3000 (I)	RealSense camera	-	RGB/NIR/Depth	5 Papercut
CSMAD [9]	2018	14	88/160 (V)	RealSense, Compact Pro, Nikon P520	-	RGB/NIR/Depth	1 silicone mask
SiW-M [169]	2019	493	660/1630 (V)	Logitech C920, Cannon EOS T6	-	RGB	1 Print, 1 Replay, 5 3D mask, 3 Makeup, 3 Partial
WFFD [138]	2019	745	2,300/2,300 (I)	Various cameras	-	RGB	1 wax figure face
CelebA-Spoof [259]	2020	10,177	202,559/475,408 (I)	Various cameras/ 20 smartphones, 2 webcams, 2 tablets	PC, Phones, Tablets,	RGB	3 Print, 3 Replay, 1 3D mask, 3 Paper Cut
PADISI-Face [194]	2021	360	1,105/924 (V)	Self-built sensor	-	RGB/NIR	1 Print, 4 Mask, 1 Makeup, 1 Tattoo, 2 Partial
CRMA [84]	2021	47	423/12,690 (V)	Various webcams/ iPad Pro, Galaxy Tab S6, Surface Pro 6	iPad Pro, Galaxy Tab S6, Surface Pro 6	RGB	1 Print, 3 Replay, **1 Real mask**

Table 2.2.: The summary of face PAD datasets, including the CRMA dataset presented in this thesis. #Sub refers to the number of subjects and BF is bona fide.

images captured by IrisAccess LG4000 camera, 2) AD100 consisting of 900 images captured by risGuard AD100 camera. Both the training and the test set are divided equally into no lens (bona fide), soft lens (bona fide), and textured lens (attack) classes.

NDCLD-2015 [142]: The 7300 images in the NDCLD-2015 [142] were captured by two sensors, IrisGuard AD100 and IrisAccess LG4000 under MIR illumination and controlled environments. The NDCLD15 contains iris images wearing no lenses, soft lenses, and textured lenses.

IIIT-D CLI [151, 239]: IIIT-D CLI dataset contains 6570 iris images of 101 subjects with left and right eyes. For each identity, three types of images were captured: 1) no lens, 2) soft lens, and 3) textured lens. Iris images are divided into two sets based on captured sensors: 1) Cogent dual iris sensor and 2) VistaFA2E single iris sensor.

LivDet-Iris 2017 dataset [243]: Though the new edition LivDet-Iris competition was held in 2020, we still evaluate the algorithms in datasets provided by LivDet-Iris 2017 in this thesis for several reasons: 1) No official training data was announced in the LivDet-Iris 2020 because the organizers encouraged the participants to use all available data (both publicly and proprietary) to enhance the effectiveness and robustness. 2) The test data is not publicly available. Consequently, to make a fair comparison with state-of-the-art

Database		Spectrum	# Training	# Testing	Type of Iris Images
NDCLD-2015 [142]		NIR	6,000	1,300	BF, soft and textured lens
NDCLD-2013 [143]	LG4000	NIR	3,000	1,200	BF, soft and textured lens
	AD100	NIR	600	300	BF, soft and textured lens
IIIT-D CLI [151, 239]	Cognet	NIR	1,723	1,785	BF, soft and textured lens
	Vista	NIR	1,523	1,553	BF, soft and textured lens
LivDet-Iris 2017 [243]	Clarkson (cross-PAD)	NIR	4937	3158	BF, textured lens, printouts
	Notre Dame (cross-PA)	NIR	1,200	3,600	BF, textured lenses
	IIITD-WVU (cross-DB)	NIR	6,250	4,209	BF, textured lenses, printouts, lens printouts
PAVID [190]		VIS	180 (V)	612 (V)	BF, replay

Table 2.3.: Characteristics of the main publicly available iris PAD datasets. All datasets have the training and test sets based on their own experimental setting in the related papers. The Warsaw dataset in the Iris-LivDet-2017 competition is no longer publicly available. BF: bona fide, VIS: visible light, NIR: Near-Infrared light. (V) indicates the video data format, others are image data format.

algorithms on equivalent data, we use LivDet-Iris 2017 datasets to restrict the evaluation factors to the algorithm itself rather than the data. 3) The LivDet-Iris 2017 competition datasets are still valuable due to the challenging cross-PA and cross-dataset scenario settings. The Clarkson and Notre Dame dataset are designed for cross-PA scenarios, while the IIIT-WVU dataset is designed for a cross-dataset evaluation due to the different sensors and acquisition environments. The Clarkson testing set includes additional unknown visible-light image printouts and unknown textured lenses (unknown pattern). Moreover, Notre Dame focused on unknown textured lenses. However, the Warsaw dataset is no longer publicly available.

Presentation Attack Video Iris Database (PAVID) [190]: PAVID is the video iris dataset collected using smartphones (Nokia Lumia 1020 and iPhone 5S) in the visible spectrum. PAVID contains 304 bona fide videos and 608 replay attack videos across 76 subjects. Moreover, PAVID was divided into three sets in the official protocol: training set including 180 videos, development set including 120 videos, and testing set including 608 videos. The development set defined in [190] was used only for determining the filter kernel of the Laplacian pyramid in [190], not for computing the decision threshold. Therefore, we omit the development set in our experiments.

2.5. Summary

This chapter introduced the general biometric systems along with the PAD systems. The main component of PAD was discussed in Section 2.1. Then, the metrics for PAD performance and biometric fairness assessment were presented. Furthermore, several fundamental layers of neural network architectures, data augmentation techniques, and attention mechanisms were introduced. This chapter also provided an overview of the popularly used deep PAD architectures along with training loss functions and supervision strategies. A summarized description of the mainstream evaluation face and iris PAD datasets were presented. The next chapters will deeply explore and address the research questions posed in Chapter1.

3. Fairness in Face PAD

The previous chapter provided essential background knowledge for biometric recognition systems and their components, along with the performance evaluation metrics, including those measuring the performance of PAD and the vulnerability of FR systems. This chapter targets the under-explored gaps in analysing, representing, and enhancing fairness in face PAD, as responses to *RQ1* and *RQ2*. This chapter is based on the [72, 81, 86].

3.1. Introduction

Data-driven approaches, including FR algorithms, are known to be unfair between certain demographic and non-demographic groups [8, 64, 69, 208, 216, 218]. The learnable models are heavily impacted by data-induced biases because they tend to optimize the objectives toward the majority group, i.e. data represented with more samples in training datasets. This commonly results in less optimized performance for minority groups, leading to unfair decisions. Fairness assessment and enhancement in biometric systems have gained increasing attention from the research community and the general public. For example, many studies have investigated fairness in FR [3, 8, 64, 218] and face quality [197, 216].

Most studies on biometric fairness have only concentrated on demographic covariants, especially gender and race. Only a few studies [216, 218] investigated the impact of other variations, such as appearance traits, on biometric fairness. Moreover, the fairness of PAD has not been investigated, except for very limited studies on iris PAD [81] and [4] face PAD. Fang *et al.*addressed the gender fairness in iris PADs and the experimental results pointed out that female users were significantly less protected by the PAD in comparison to males [81]. Alshareef *et al.*[4] considered the gender fairness in face PAD by using ResNet50 [108] and VGG16 [207] on a limited PAD data. Both studies focused only on the gender fairness assessment and with very limited evaluation data, both in terms of size and diversity. The main reason is that the majority of publicly available PAD datasets do not contain information regarding demographic and non-demographic attributes, making it impossible to assess fairness, let alone enhance potential unfairnesses.

Moreover, the fairness of PADs in both studies [4, 81] were measured by differential performance and outcome of PAD, i.e. no fairness metrics were applied. To date, only two very recent definitions of fairness in FR were proposed, Inequality Rate (IR) [98] and Fairness Discrepancy Rate (FDR) [64]. IR [98] takes ratio differences between minimum and maximum FR performance per group. However, IR has two drawbacks: 1) IR has no theoretical upper bound due to its multiplicative nature and exponential weights, 2) IR might be incomputable due to its ratio property, i.e. when minimum FR performance (denominator) for any group is zero. FDR [64] considers the maximum difference FR performance between any two groups based on a decision threshold calculated on all groups. In our work, we adapt FDR to represent fairness in PAD performance. However, FDR does not take absolute performance into account and thus might consider a fair but low-performing PAD "better" than an unfair PAD that performs close to perfect.

To target these under-explored gaps, this chapter analyses, represents, and enhances fairness in face PAD. In the efforts seeking to answer *RQ1* and *RQ2*, this chapter presents the following contributions:

- To enable answering *RQ1* and *RQ2*, this chapter presents a Combined Attribute Annotated PAD Dataset by combining six publicly available PAD datasets including highly diverse PAs, in which seven human-annotated attribute labels covering demographic and non-demographic attributes are provided. Moreover, this chapter adapts the fairness metric to face PAD and introduces a novel metric, named Accuracy Balanced Fairness (ABF), to represent both the PAD fairness and the absolute PAD performance simultaneously.

- As a response to *RQ1*, this chapter assesses the fairness of face PADs from two aspects, the nature of the training data and the Operational Decision Threshold Assignment (ODTA) based on data of different groups.

- As a response to *RQ2*, this chapter proposes a simple data augmentation solution, named FairSWAP, to disrupt the identity/semantic information, boosting PAD performance and fairness in most cases, as will be demonstrated in a detailed and diverse experiment.

The rest of this chapter is organized as follows: Section 3.2 presents the essential background information regarding fairness in automated decision systems and PAD. Section 3.3 introduces the presented Combined Attribute Annotated Presentation Attack Detection Dataset (CAAD-PAD), including descriptions of each used PAD dataset, human-annotation criteria, and experimental protocols designed to enable fairness analyses. Section 3.4 focuses on setups for fairness assessment and provides the description of used

PAD algorithms, evaluation metrics including our proposed ABF, and implementation details. Section 3.5 introduces the proposed FairSWAP for fairness enhancement and the corresponding implementation details. Section 3.6 and 3.7 discuss the results of fairness assessment and enhancement of face PADs, respectively. A summary is presented in Section 3.9.

3.2. Related work

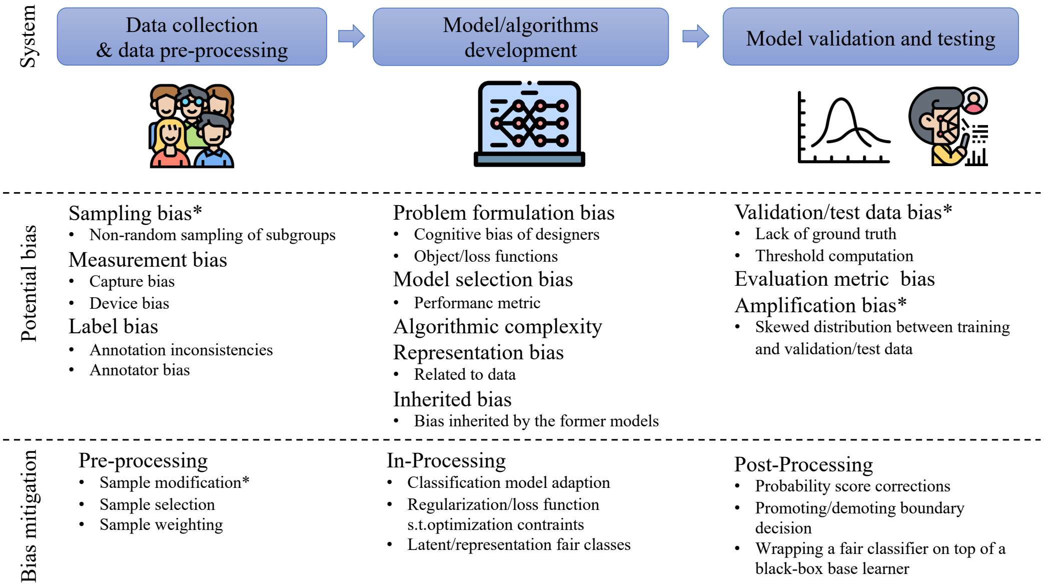

System
Data collection & data pre-processing
Model/algorithms development
Model validation and testing
Potential bias
Sampling bias*
• Non-random sampling of subgroups
Measurement bias
• Capture bias
• Device bias
Label bias
• Annotation inconsistencies
• Annotator bias
Problem formulation bias
• Cognitive bias of designers
• Object/loss functions
Model selection bias
• Performanc metric
Algorithmic complexity
Representation bias
• Related to data
Inherited bias
• Bias inherited by the former models
Validation/test data bias*
• Lack of ground truth
• Threshold computation
Evaluation metric bias
Amplification bias*
• Skewed distribution between training and validation/test data
Bias mitigation
Pre-processing
• Sample modification*
• Sample selection
• Sample weighting
In-Processing
• Classification model adaption
• Regularization/loss function s.t.optimization contraints
• Latent/representation fair classes
Post-Processing
• Probability score corrections
• Promoting/demoting boundary decision
• Wrapping a fair classifier on top of a black-box base learner

Figure 3.1.: Aspects effecting fairness in automated decision systems. $*$ indicates the addressed aspects in this work, where sampling bias, validation/test data bias, and amplification bias are relevant to the nature of training data, the ODTA on different data groups, and the training and test on group-disjoint data, respectively.

3.2.1. Fairness in automated decision systems

Biometric systems, as automated decision-making systems, has been widely deployed in recent decades. Building an automated decision-making system requires the following steps: data collection and pre-processing, model/algorithm development, model validation and testing, as illustrated in 3.1. Data, such as face and iris traits, are collected by sensors and pre-processed as input for learnable models. Model/algorithm development aims to summarize the pattern of biometric traits via supervised/unsupervised learning. The developed model is then evaluated with respect to certain metrics, such as accuracy and equal error rate (EER), among others. As suggested in [64, 202], the potential unfairness triggers may exist at every stage of an automated system, as listed in Figure 3.1 (* refers to aspects addressed in this work). Potential unfairnesses occur during the data collection phase, such as sampling bias causing data imbalance, measurement bias related to the capture environments and sensors, and label bias attributed to annotators. Similarly, model/algorithm development could be human biased (e.g., problem formulation or objective function definition by human designer) or statistically biased (e.g., performance metric based model selection or bias inherited from the training data or pre-processing), as argued in [46, 60, 202]. Lastly, fairness issues could arise in the model validation and testing given an account of potential validation/test bias, evaluation metric bias or amplification bias [202]. Amplification bias occurs when the validation/test data is skewed in comparison to the training data distribution, which motivates our experimental protocol design in Section 3.3.3. In automated biometric systems, a threshold commonly computed from the validation (development) set is necessary to make a final decision. The unfair distribution of validation data could be transmitted to the threshold computation, which is an issue analysed in this work. Schwartz *et al.*[202] suggested that the dataset, the evaluation, and the human factor are the three critical challenges in machine learning fairness. Fairness issues caused by human factors are much more complex and multi-faceted including societal and historical aspects. Therefore, this work focuses on studying the dataset- and ODTA -related fairness concerns and proposing a fairness enhancement solution on the data pre-processing level.

3.2.2. Fairness in face presentation attack detection

Recently, Drozdowski *et al.*[69] presented a comprehensive summary of the existing literature on fairness assessment and enhancement of biometric systems. This survey found that a majority of existing studies has conducted experiments using face traits and has concentrated only on recognition algorithms, while the fairness of PADs has barely been investigated. Up to now, far too little attention has been paid to the fairness of face PAD systems. To the best of our knowledge, only one study [81] explored the gender fairness of iris PAD systems and one work [4] considered the gender fairness of face PAD by using ResNet50 [108] and VGG16 [207] on one face PAD dataset. The fairness of demographic attributes (e.g., gender, age, race) and soft-biometric attributes (e.g., beards, hair, makeup, accessories) of face PAD systems has been extremely understudied.

The main possible reason is the insufficient face PAD data and the lack of such attribute labels. For instance, most of the existing PAD datasets, e.g., OULU-NPU [15], CASIA-MFSD [262], MSU-MFS [234], Idiap Replay-Attack [44], were collected in laboratories from insufficient subjects (details can be found in Table 3.1) and thus leading to the limited variations in soft-biometric characteristics. A large-scale face PAD dataset, CelebA-spoof [260], provided some additional soft-biometric labels. However, after a comprehensive analysis on the data distribution, we found that data in attribute groups in CelebA-sooof is extremely unbalanced over bona fide and attack samples, making it impossible to design proper experimental protocols for PAD fairness analyses. For example, subjects with eyeglasses, makeup, or bangs are all bona fide samples. The same issue exists in the PADISI-Face [195] dataset. To address this issue, this work combined six publicly available face PAD datasets consisting of print, replay, 3D mask and wax figure attacks, and provided seven human-annotated attribute labels to enable PAD fairness studies.

In addition to the lack of appropriate labeled data, there is no standard criteria to assess the fairness of developed systems. Fang *et al.* [81] studied the demographic bias by adapting and reporting the differential performance and differential outcome as suggested for verification performance in [115]. Differential performance as in [81] measures the difference in the bona fide or attack decision distribution between specific attribute groups independently of any decision threshold, while differential outcome describes the difference in APCER or BPCER rates between different demographic groups relative to a group-specific decision threshold. Alshareef *et al.* [4] measured the demographic bias and fairness of PAD solutions by observing the difference in PAD performance, such as Area Under the Receiver operating characteristic curve (ROC-AUC), accuracy, EER, and APCER/BPCER. Both studies assessed fairness either using a group-specific decision thresholds or being independent of any decision threshold

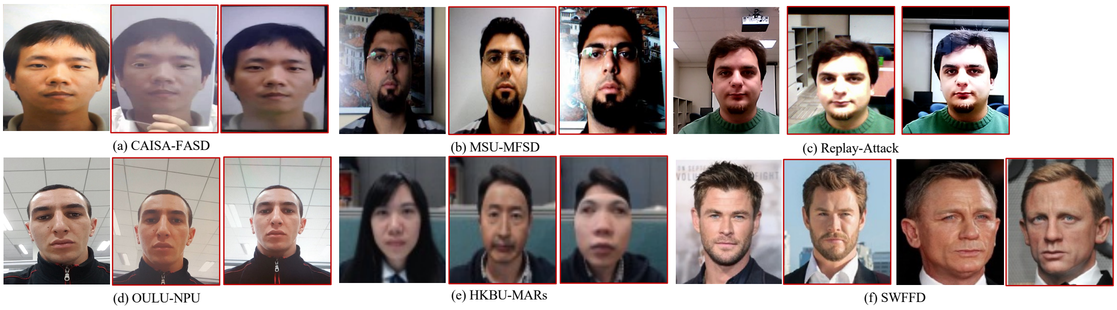
However, using such group-specific thresholds is not fair for different groups, as mentioned in [64], and measuring fairness independently from thresholds is not realistic. Such problems was raised and explored in FRs in [154]. Our work also assess the fairness associated with ODTA of face PADs by exploring the PAD performance under different attribute group decision thresholds (corresponds to the marked points of the model validation/testing phases in the Figure 3.1). Furthermore, to bridge the fairness measurement gaps, Pereira and Marcel [64] introduced Fairness Discrepancy Rate to evaluate and compare the fairness between biometric verification systems. Unlike measuring the fairness by reporting differential performance [4, 69, 81], FDR assessed the trade-off between the model performances by assuming a single 'fair' decision threshold for all demographic groups. Therefore, FDR is adapted in this study to assess the fairness of face PADs (details in Section 3.4.2). However, FDR does not consider the absolute performance, thus a completely fair but low performing PAD will be considered 'better' than a slightly unfair

PAD that performs perfectly, according to FDR. To further link the PAD performance and fairness, we propose the ABF.

3.3. Combined Attribute Annotated PAD Dataset

Extensive research [3, 8, 64, 69, 218] has shown that recognition systems exhibit bias, that is, subjects in a certain demographic or non-demographic groups are more accurately recognized than other groups. However, most of these research efforts focus on the exploration [8, 64, 218, 225] or mitigation [25, 95, 209, 219] of the bias only in FR algorithms. There has been no detailed investigation of demographic and non-demographic bias in face PAD systems. Besides the contemporary nature of biometric bias studies, a possible reason is the lack of sufficient PAD data with soft-biometric labels. To this end, we alleviate this issue by combining six face PAD datasets and providing publicly released corespondent annotations covering demographic and non-demographic attributes, named Combined Attribute Annotated PAD Dataset (CAAD-PAD). The detailed description of each selected PAD dataset, the criteria of annotations, the distribution of CAAD-PAD and the experimental protocols for fairness assessment are presented in the following Section 3.3.1, 3.3.2, and 3.3.3, respectively.

3.3.1. Datasets



(a) CAISA-FASD

(b) MSU-MFSD

(c) Replay-Attack

(d) OULU-NPU

(e) HKBU-MARs

(f) SWFFD

Figure 3.2.: Samples from the datasets used to built our CAAD-PAD (attacks in red frame).

The selected six face PAD datasets (to build our CAAD-PAD) are presented in details in this section and Table 3.1 summarizes the information of these PAD datasets, as well as other two conventional PAD databases which contain attribute labels. CelebA-Spoof dataset [260] involving rich annotations (as listed in Table 3.1) is considered unsuitable for

Dataset	Year	# BF/attack	# Sub	Attack types	Attribute label
CASIA-FASD [262]	2012	150/450 (V)	50	1 Print, 1 Replay	No
Replay-Attack [44]	2012	200/1,000 (V)	50	1 Print, 2 Replay	No
MSU-MFSD [234]	2015	70/210 (V)	35	1 Print, 2 Replay	No
HKBU-MARs [166]	2016	120/60 (V)	12	2 3D masks	No
OULU-NPU [15]	2017	1,980/3,960 (V)	55	2 Print,2 Replay	No
SWFFD [135]	2019	2,300/2,300 (I)	745	1 3D face	No
CelebA-Spoof [260]	2020	184,407/ 377,168(I)	10,177	3 Print, 3 Replay, 1 3D, 3 Paper Cut	Yes (40)
PADISI-Face [195]	2021	1,105/924 (V)	360	1 Print, 4 Mask, 1 Makeup, 1 Tattoo, 2 Partial	Yes (4)
CAAD-PAD (our)	2022	2,510/5,680(V) 2,300/2,300(I)	947	3 Print, 2 Replay, 2 3D masks, 1 wax face	Yes (7)

Table 3.1.: Summary of the main existing face PAD datasets. "V" and "I" refers to video and image sample, respectively. "BF" indicates the bona fide samples. The number following the "Yes" is the number of attribute label types.CAAD-PAD (our) combines CASIA-FASD [262], Replay-Attack [44], MSU-MFSD [234], HKBU-MARs [166], OULU-NPU [15] and SWFFD [135]. Despite the rich annotations of CelebA-Spoof [260], the provided attributes are extremely unbalanced over bona fide and attack, making it impossible to design proper protocols for fairness assessment. The PADISI-Face [195] has the same issue and is of a much smaller scale. These issues are discussed in details in Section 3.3 and motivates the need for our CAAD-PAD dataset.

the goal of this work work due to the extreme imbalance of its data making it impossible to design proper experimental protocols for fairness study. For example, subjects with attributes, such as eyeglasses, makeup and bangs, are all bona fide samples in the CelebA-Spoof dataset. The similar issue exists in the PADISI-Face [195] dataset. To enable fairness assessment of face PAD, we selected the following publicly available face PAD datasets (samples in Figure 3.2): CASIA-FASD [262], Replay-Attack [44], MSU-MFSD [234], HKBU-MARs [166], OULU-NPU [15] and SWFFD [135], to be composed into our CAAD-PAD. This combination of datasets are chosen as: 1) they consist of diverse PAs including print and replay attacks, 3D mask attacks, and wax figure face attacks. 2) CASIA-FASD [262], Replay-Attack [44], MSU-MFSD [234], and OULU-NPU [15] are widely used in generalization studies [83, 168, 200, 204] which is the main problem in PAD. 3) HKBU-MARs [166] and SWFFD [135] target realistic mask attack problem which is one of the practical PA problems in real-world applications. The description of each selected dataset are provided in the following:

CASIA-FASD [262] dataset covers a diverse potential attack types in real world including warped photo, cut photo, and video attacks. CASIA-FASD includes 600 videos across 50 subjects where each subject has 12 videos (three bona fide and nine attack). The videos consist of low, normal and high quality, where videos of low and normal quality have a resolution of 640×480, and high quality videos have a resolution of 1280×720 pixels.

Idiap Replay-Attack [44] dataset contains 1,200 videos (200 bona fide and 1,000 attack) of 50 subjects. The bona fide and attack videos were acquired in two sessions: controlled and adverse where the background and illumination conditions are not uniform. The attack videos were captured under two modes: hand-based attack (i.e. the attack device is held by human hand) and fixed-support attack (i.e. the attack device is placed on a fixed support).

MSU-MFSD [234] dataset consists of 440 videos (110 bona fide and 330 attack videos) across two attack types, video replay and printed photo, of 55 subjects.

OULU-NPU [15] dataset was captured under the realistic mobile authentication scenarios, where a total of 5,940 videos (1,980 bona fides and 3,960 attacks) of 55 subjects were captured under various illumination and backgrounds by six different mobile phones.

HKBU-MARs [166] is a 3D mask PAD dataset where the 3D masks were generated by two different techniques. First is the ThatsMyFace mask which leverages 3D reconstruction and 3D printing techniques to generate the facial mask. The advantage of ThatsMyFace Mask is that the mask can be easily generated with a single customized image, but the skin texture is not well restored due to the defects of 3D printing technique. The second technique is the REAL-F mask[1] which provides high quality appearance and looks very similar to a bona fide face as the skin texture, blood vessels of the eyes and iris can be reconstructed. This database contains 180 videos (120 bona fides and 60 attacks) of 12 subjects. An example of bona fide sample, a ThatsMyFace attack and a REAL-F mask are shown in Figure 3.2 (e).

Single Wax Figure Face Database (SWFFD) [135] dataset provides wax figure faces as realistic 3D PAs, serving as complementary of the above face PAD datasets including replay, print, and other 3D attacks. WFFD consists of 2,200 images with both bona fide and wax figure attack faces (totally 4,400 faces of 745 subjects) with a high diversity from online collections. Examples of image of SWFFD [135] are shown in Figure 3.2 (f).

However, these PAD database contains only PAD labels and lack any face attribute labels. To enabling further PAD fairness assessment in demographic and non-demographic groups, we provide seven soft-biometric annotations as discussed in the next section.

[1] `https://real-f.jp`

3.3.2. Face attribute annotation criteria

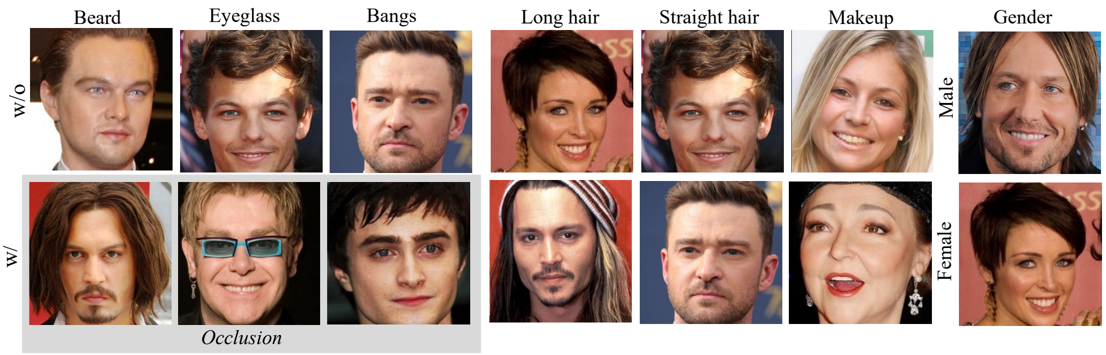


Figure 3.3.: Samples of each considered attribute. Samples are labeled as occlusion if they have beard, eyeglass, bangs, or a combination of them. Samples without beard, bangs, and not wearing eye glass are labeled as non-occlusion. The other attributes are individually labeled.

We manually annotated the CAAD-PAD dataset under the following criterion (as shown in Figure 3.3):

- **Gender**: Gender metric in our case is determined based on the human perception of the gender from the face image. In this work, a subject is categorized into either male or female groups based on the majority decision of five annotators. These decisions might have been also influenced by previous knowledge of the subjects as some datasets included celebrities.
- **Beard**: A subject with no visible hair coverage around the mouth or shaved with only light hair roots is labeled as no beard.
- **Eyeglasses**: A subject wearing eyeglasses in the face area is annotated as wearing eyeglasses, regardless of types, shapes and colors of eyeglasses. Note that the eyeglasses in the area above the forehead is labeled as no eyeglasses, as eyeglasses in such position will be removed after the face detection and cropping.
- **Bangs**: A subject with hair covering more than 15% of the forehead is considered to have bangs.
- **Makeup**: A subject with noticeable colors of lipstick and eye shadow is categorized as makeup group.

- **Long/short hair**: A subject with hair beyond shoulder is considered to long hair group.
- **Curly/Straight hair**: A subject with noticeable waves of hair is categorized into curly hair group. Such as gender, all the annotations were based on the majority decision of five annotators.

Overall, the combined CAAD-PAD contains 8,190 videos and 4,600 images across 947 subjects, covering the seven binary attributes.

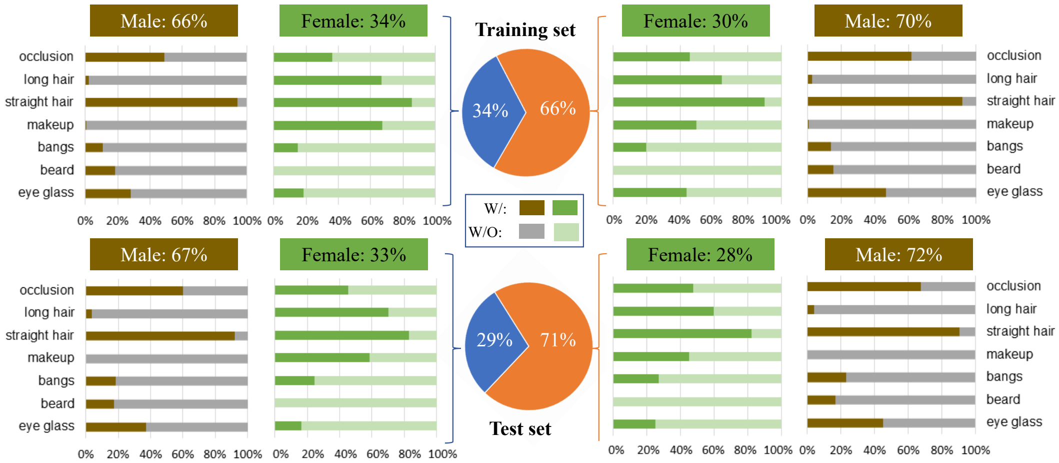


Figure 3.4.: Data distribution of the training and test set of CAAD-PAD. Most of the attributes are understandably not well-distributed over genders (but well distributed over bona fide and attacks), e.g. makeup of male sample, beard of female sample, long hair of male sample. Samples with occlusion and without occlusion are well-distributed over genders.

3.3.3. Experimental protocols

This section first presents an in-depth analysis of the data distribution of the CAAD-PAD dataset (as shown in Figure 3.4). The training set is a combined set of training and development subsets from six used datasets. Similarly, the test set of CAAD-PAD is a combined set of test subsets from each selected dataset. Therefore, the subjects in the training and test set are disjoint. The detailed information is provided as follows:

- **Training set**: As most selected datasets have pre-defined and identity-disjoint training, development, and test sets, the training set of CAAD-PAD is composed of samples from the original training and development subsets. Only HKBU-MARs [166] does not provide the pre-split subsets. As a result, we manually selected the subjects to form the identity-disjoint training and test sets of HKBU-MARs by considering the subject attribute annotations with a subject ratio of 8:2 (train to test). The ratio of bona fide and attack samples is 1:1.9, and the ratio of female and male samples is 1:2.1. Samples with occlusion and without occlusion are well-distributed over genders (close to 1:1). Although samples of several facial attributes (e.g., makeup, beard, long hair) are not well-distributed over genders, such samples in CAAD-PAD are very well balanced over bona fide and attack, unlike CelebA-Spoof [260] and PADISI-Face [195] datasets.

- **Test set**: The test set consists of the data samples from the pre-defined test subsets in each dataset (except HKBU-MARs [166] is identity-disjoint split by us). Similar distribution to the training set is observed in Figure 3.4.

To assess the PAD fairness on different groups, the following protocols are designed to target three aspects: gender, occlusion, and facial attributes, as well as the effect of the training data on fairness. The designed protocols are as follows:

- **Protocol-1**: Protocol-1 targets the fairness of PAD performance across gender groups. Protocol-1 consists of three sub-protocols:
  - protocol 1.1: PAD solution is trained on the above fused training set (including female and male) and tested separately on female test set and male test set.
  - protocol 1.2: This protocol studies gender bias when male data is unseen in the training phase. Thus, PAD model is trained only on female data in the training set and tested separately on female and male data in test set.
  - protocol 1.3: contrary to to protocol 1.2, PAD model in this protocol is trained only on male data in the training set, and tested separately on female and male samples in the test set to study gender bias case where female data is unseen during the training phase.
- **Protocol-2**: Protocol-2 targets the fairness of PAD performance across occlusion groups (occlusion and non-occlusion).
  - protocol 2.1: PAD model is trained on the fused training set (including samples with and without occlusion) and tested separately on occluded and non-occluded samples.

  - protocol 2.2: This protocol explores fairness when the model only learns from occluded samples. Therefore, the PAD model is trained on samples with occlusion and tested separately on samples with and without occlusion.
  - protocol 2.3: Conversely, PAD solution is trained only on non-occluded samples in the training set and tested separately on samples with and without occlusion in the test set.

- **Protocol-3**: Protocol-3 explores the fairness of PAD performance across the remaining attribute groups (with/without eyeglass, beard, bangs, long/short hair, straight/short hair), respectively. As some attributes are unbalanced over genders, such as makeup and beard, we only train PAD models on the fused training set, and test on each attribute group separately, unlike the extended protocols over gender and occlusion.

3.4. Fairness assessment

This section first introduces the four PAD solutions used as the bases of our fairness assessment. Then, the metrics for PAD performance and fairness evaluation are presented. In addition, we refine the used fairness metric and introduce a novel metric that links the PAD biometric fairness with the worst PAD cases among all groups, as will be motivated in details. Lastly, the implementation details are provided to insure reproducibility.

3.4.1. PAD algorithms

To assess the fairness of PAD solutions, we adopt four well-established and diverse PAD solutions ranging from the hand-crafted feature based to deep learning based methods: LBP-MLP [63], ResNet50 [108], DeepPixBis [94], LMFD [83]. **LBP-MLP [63]** Considering that LBP is a widely used hand-crafted feature in earlier PAD studies [16, 63, 173], we use LBP features to investigate the fairness of face PAD following the highly influential work in [16, 173]. Two LBP feature vectors are extracted from each image in RGB and HSV color space, separately. Then, two feature vectors are concatenated into one vector of dimension 60×1. To detect an input image is bona fide or attack, the concatenated feature vector is fed to a simple Multi-Layer Perceptron (MLP) classifier consisting of only two fully-connected layer.
ResNet50 [108] Residual learning framework was firstly proposed in [108] aiming to ease and stabilise the training of networks, since computation complexity increases as network grows. Considering that ResNet50 [108] was applied as a backbone neural architecture in

many PAD methods [137, 247, 260] and achieved good PAD performance, we use it in our experiments to further assess its fairness.

DeepPixBis [94] DeepPixBis [94] is the first work to adopt pixel-wise binary supervision to enhance the PAD performance. The pixel-wise binary label on the output maps forces the network to learn fine-grained representation from different pixels/patches. DeepPixBis [94] showed the good PAD performance under intra-dataset scenario and thus we adopt in our PAD fairness assessment.

LMFD-PAD [83] LMFD-PAD [83] presented a dual-stream PAD framework, in which one stream was used to learn features in the frequency domain and the other stream was used to learn features in spatial color space. The benefit of this architecture is that feature in frequency domain is less influenced by data capture devices and environment information, resulting in one of the current top performing face PADs in the literature. We used this PAD solution by considering its high PAD generalizability under the cross-dataset evaluation.

Implementation details of all solutions are further discussed in Section 3.4.3.

3.4.2. Evaluation metrics

To measure the performance and the fairness of the PAD algorithms, we adopt the widely used PAD metrics defined in the standardized ISO/IEC 30107-3 [10] and a fairness measurement metric for FR systems introduced recently in [64], which makes the bases of our proposed PAD fairness metric.

PAD performance metric: Following the definitions in ISO/IEC 30107-3 [10], Attack Presentation Classification Error Rate (APCER) and Bona Fide presentation classification error rate (BPCER) are used. APCER refers to the proportion of attack presentations incorrectly classified as bona fide presentations and BPCER refers to the proportion of bona fide samples misclassified as attack samples. To cover different operation points, we report the 1-BPCER value at six different fixed APCER values (0.5%, 1%, 5%, 10%, 15%, 20%) in plots. To provide a clear and straightforward comparison, we plot these values in Figures (as shown in Figure 3.6, 3.10, 3.8, 3.11). In addition, Equal Error Rate (EER), the APCER value when APCER and BPCER are equal, is also reported to measure the overall PAD performance in Tables 3.2, 3.4, 3.6.

Biometric fairness metric: To explore the fairness of PAD performance on different attribute groups, we adapt the Fairness Discrepancy Rate. FDR is proposed in [64] to assess fairness of a biometric verification systems by leveraging a single decision threshold from all test groups. Their experimental assessment based on pre-built fair and unfair systems

[64] illustrated that FDR can better represent the fairness of algorithms in comparison to only using ROC/DET curves. As a result, FDR we utilize to verify the fairness of PAD solutions. A PAD system is considered fair if different attribute groups share the same BPCER and BPCER for a given decision threshold τ, where $\tau = APCER_x$ from all groups. Based on this theorization, the FDR can be calculated as follows:

$$A(\tau) = max(|APCER^{d_i}(\tau) - APCER^{d_j}(\tau)|), \quad \forall d_i, d_j \in D \tag{3.1}$$

$$B(\tau) = max(|BPCER^{d_i}(\tau) - BPCER^{d_j}(\tau)|), \quad \forall d_i, d_j \in D \tag{3.2}$$

$$FDR(\tau) = 1 - (\alpha A(\tau) + (1 - \alpha)B(\tau)) \tag{3.3}$$

where D is a set of attribute groups $D = \{d_1, d_2, ..., d_n\}$, α is a hyper-parameter that defines the importance of misclassified attacks (i.e. APCER). The value of FDR varies from 0 (maximum discrepancy) to 1 (minimum discrepancy). Therefore, FDR has smaller-is-worse semantics. We plot the FDR values under different decision thresholds $\tau = APCER_x$ where x varies from 0.005 to 0.2. As stated in [64], the FDR value of a fair system is not sensitive to the variation of τ, which reflected in the plot is that the FDR curve should be straight and in a higher position. The α is set to 0.5 in our experiments as suggested in [64], but can be chosen differently if a certain application considers that the fairness in one of the errors (BPCER or BPCER) is more important.

Accuracy Balanced Fairness: To explore the fairness of PAD performance on different attribute groups, we adapt the FDR. However, FDR does not consider absolute performance, thus a completely fair but low-performing PAD will be considered "better" than a very slightly unfair PAD that performs perfectly, according to FDR. To further associate the PAD performance and biometric fairness, we propose the Accuracy Balanced Fairness metric. Given a decision threshold τ, where $\tau = APCER_x$ from all groups, ABF is formulated as follows:

$$A(\tau) = \frac{max(|APCER^{d_i}(\tau) - APCER^{d_j}(\tau)|)}{1 - max_D(APCER(\tau))}, \quad \forall d_i, d_j \in D \tag{3.4}$$

$$B(\tau) = \frac{max(|BPCER^{d_i}(\tau) - BPCER^{d_j}(\tau)|)}{1 - max_D(BPCER(\tau))}, \quad \forall d_i, d_j \in D \tag{3.5}$$

$$ABF(\tau) = 1 - (\alpha A(\tau) + (1 - \alpha)B(\tau)) \tag{3.6}$$

Numerators of $A(\tau)$ and $B(\tau)$ are the same as the FDR metrics, while denominators $1 - max_D(APCER(\tau))$ and $1 - max_D(BPCER(\tau)$ are used to weight the discrepancy by considering the worst PAD cases in a set of attribute groups. As a result, a higher ABF value is linked to a smaller fairness discrepancy and a higher PAD performance. Similar to FDR values, we compute the ABF values under different decision thresholds $\tau = APCER_x$ where x varies from 0.005 to 0.2 and the area under the ABF is provided for overall quantitative comparison.

3.4.3. Implementation details

In our experiments, we sampled 20 frames at equal intervals from each video (when the data format is video) for training and testing, following the common practice of frame sampling in PAD [83, 94]. For each frame, the face was detected and cropped by Multi-task cascaded convolutional neural network (MTCNN) [256] and resized to $224 \times 224 \times 3$ pixels [83, 94]. For training an LBP-MLP model, the Adam optimizer was used with an initial learning rate of 10^{-1} and binary cross-entropy loss was used to supervise the MLP training. For training the DeepPixBis [94] and LMFD-PAD [83], we followed the implementation setups described in their respective works, including horizontal flip, rotation, and color jitter. For training a ResNet model, we use the same training settings for DeepPixBis [94] with cross-entropy loss only. To address the unequal distribution of samples between attacks and bona fide, we simply oversample the minority to make the ratio of bona fide and attack data close to 1:1 following the common practice [83, 94, 153, 199]. To further reduce overfitting, we also utilized the early stopping technique with the maximum epochs of 100 and the patience of 20 epochs for ResNet [94], DeepPixBis [94], and LMFD-PAD [83]. The maximum training epoch is experimentally set to 1000 for LBP-MLP [63].

3.5. Fairness enhancement: FairSWAP

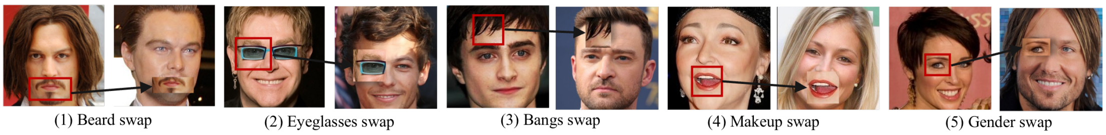


Figure 3.5.: FairSWAP: patches are swapped between training samples from different attribute groups, aiming to disorder the identity and appearance features.

In this work, four face PAD solutions (details in Section 3.4.1) are adopted to investigate the performance fairness of different groups. The experimental results reported later in Section 3.6 indicate that all assessed PAD solutions perform unfairly on different attribute groups. For example, Table 3.2 shows that the performance on the female group is worse than on the male group when the model learns from both male and female groups. Also, Table 3.6 indicates that subjects with bangs possess a worse performance than subjects without bangs. One possible reason is that face PAD algorithms may be implicitly influenced by semantic concepts related to face identity, rather than constantly focusing on discovering the discriminate features between bona fide and attacks. Motivated by this reasoning, we propose a fairness boosting cross-attribute patch swapping augmentation (FairSWAP) technique to mitigate the bias in the trained PAD solutions. The FairSWAP swaps image patches between images of different attribute groups to disrupt the demographic/semantic information and guide PAD models to learn to differentiate between attack and bona fide clues, rather than other irrelevant clues. As shown in Figure 3.5, given a training image, the attribute region of a subject, such as beard, eyeglasses, bangs, and makeup (e.g. lipstick), is extracted and overlaid on the sample of another training subject. In the case of gender groups, a randomly located (details to follow) region from an image of a female subject is swapped to an image of a male subject, and vice versa. In our work, the swapping process is simplified to randomly selecting a region from the candidate image and swapping it to the same position in another image, which acts as data augmentation. For PAD methods which utilize pixel-wise supervision (i.e. DeepPixBis [94] and LMFD-PAD [83]), the corresponding ground truth map is updated based on the swapped locations simultaneously. For the PAD methods which utilize binary supervision (i.e. LBP-MLP [63] and ResNet50 [108]), the label of the image is updated to attack if the input image after swapping contains a partial attack region, otherwise, the label of the image is unchanged. The swapping strategy is detailed as follows:

- Given a bona fide image x_1, a randomly located patch is swapped with a probability p_1. If FairSWAP is performed, another bona fide candidate image x_2 (i.e. all bona fide training images except x_1) will be randomly selected. The corresponding training label for training remains bona fide.
- Given an attack image x_1, the final image fed to models is determined by three probabilities. A probability p_2 is first used to decide whether to apply FairSWAP. If FairSWAP is performed, p_3 is then used to determine a randomly selected attack or bona fide candidate image (i.e. all training samples excluding x_1). Lastly, if a bona fide image is selected, p_4 is used to decide the partial attack size. The binary training labels remain attack, while the corresponding pixel-wise map is updated according to the swapped region.

As FairSWAP serves as a data augmentation technique and can be plugged into any training process, the image pre-processing and hyper-parameters (e.g., optimizer and learning rate) used to train the fairness enhancement model are the same as used in the fairness assessment in Section 3.4.3. The swapped patch size in FairSWAP is 64×64 in our experiments and the probabilities p_1, p_2, p_3, and p_4 are manually set to 0.3, 0.3, 0.5, and 0.5 based on empirically experiments.

3.6. Results of fairness assessment

This section presents the PAD performance across attribute groups in terms of detection EER and assesses the fairness under various ODTAs in terms of FDR and proposed ABF. To explore the dataset-induced potential bias, we adopt four PAD models described in Section 3.4.1 on different data groups following the experimental protocols presented in Section 3.3.3. The overall PAD performance is reported in terms of EER values in Table 3.2, 3.4, and 3.6. In addition, 1-BPCER values are computed based on different ODTAs and illustrated in Figure 3.6, 3.8. In the following, we discuss the results on gender groups, occlusion and non-occlusion groups, and the remaining attribute groups.

3.6.1. Fairness assessment over gender (Protocol-1)

Trained	Test	LBP [63]			ResNet50 [108]			DeepPixBis [94]			LMFD-PAD [83]		
		B	FairSwap	Impro	B	FairSwap	Impro	B	FairSwap	Impro	B	FairSwap	Impro
Fused	M	11.13	11.27	+0.14	2.54	1.86	-0.68	1.17	1.03	-0.14	1.94	1.72	-0.22
	F	**17.69**	**14.06**	-3.63	**3.00**	**2.34**	-0.66	**1.57**	**1.29**	-0.28	**2.62**	**2.25**	-0.37
M	M	12.55	12.88	+0.33	2.96	1.54	-1.42	1.32	1.45	+0.13	2.15	2.68	+0.53
	F	**16.17**	**17.12**	+0.95	**5.90**	**2.34**	-3.56	**3.50**	**2.56**	-0.94	**3.89**	**3.87**	-0.02
F	M	**19.04**	**22.48**	+3.44	**13.13**	**9.13**	-4.00	**10.95**	**11.77**	+0.82	**9.88**	**10.85**	+0.97
	F	18.67	21.98	+3.31	10.62	8.25	-2.37	7.52	8.42	+0.90	9.15	8.33	-0.82

Table 3.2.: PAD performance in terms of EER (%) on gender groups by baseline models (B) and FairSWAP. Fused, M, and F refer to the fused male and female data, male, and female data, respectively. Impro refers to percentage improvement. The results are obtained by separately training models on fused, male, and female training data of CAAD-PAD. Bold numbers indicate the highest EER values between male and female test data by each trained model.

PAD Performance with respect to gender The experiments were conducted following the designed Protocol-1 (Section 3.3.3) to assess the differential performance and outcome, and fairness on gender groups. Four PAD models LBP-MLP [63], ResNet50 [108], DeepPixBis [94], LMFD-PAD [83] were trained separately on a fused training set including males and females (Fused), males (M), and females (F) of CAAD-PAD. The results are then reported separately on male and female test data in Table 3.2. The EER values in

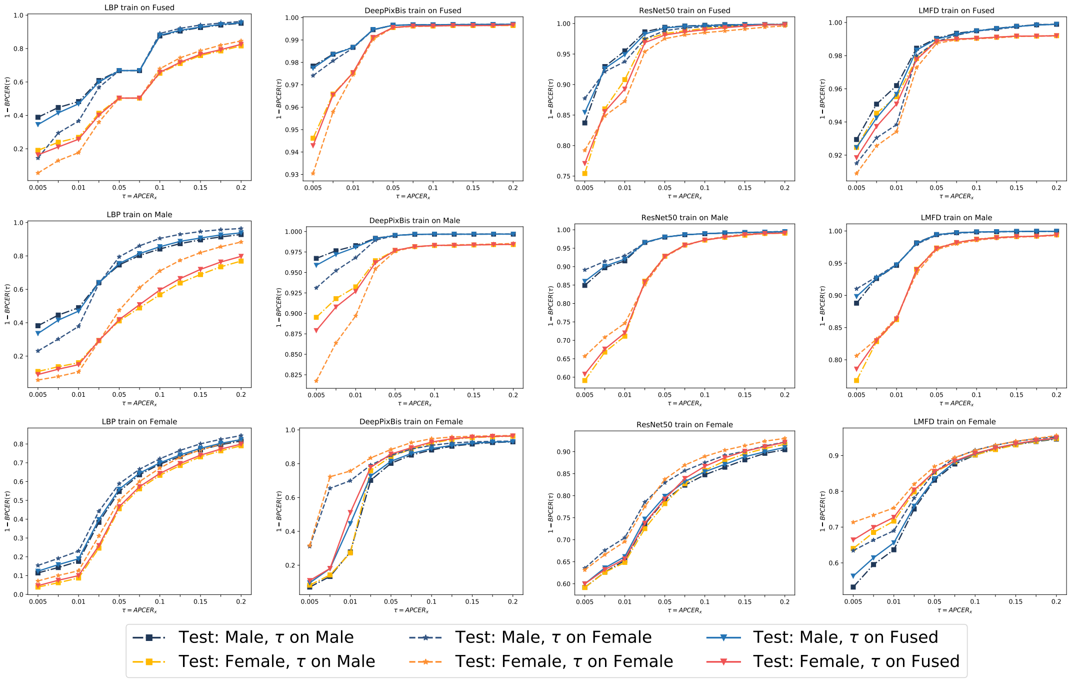


Figure 3.6.: The results of baseline models on gender groups in terms of $BPCER \ @ \ APCER_x$. To study the impact of ODTA on fairness, the threshold τ is calculated separately on fused, male, and female test set. When evaluating models trained on fused data, using a threshold computed on male samples shows a relatively higher performance than a threshold from fused or male groups.

columns 'B' refer to baseline results (results of FairSWAP will be discussed later in Section 3.7) and bold numbers refer to higher error rates between the male and female groups. In the case of models trained separately on male and female samples, the error rates are higher for the gender group that is unseen during the training phase. Such results are reasonable as it is a challenge for models to generalize on data with unknown aspects. When models are trained on the fused data, male test set obtains consistently lower EER values than females, indicating that the male group has a relatively higher protection from the PAD systems. It should be noted that the ratio of male and female samples in the fused training data is close to 1:1, achieved by data oversampling (details Section 3.4.3). Moreover, models trained on the male group achieve consistently lower EER values on both test sets than models trained on the female group. For example, when models are trained on male data, DeepPixBis [94] and LMFD-PAD [83] achieve the lowest EER

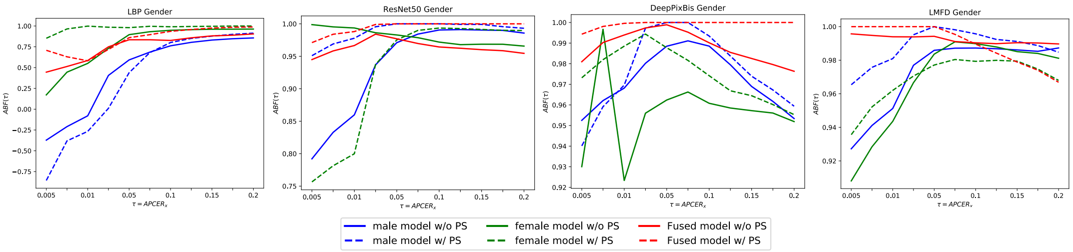


Figure 3.7.: Fairness measure in terms of ABF values by employing four face PAD models on gender groups. Solid lines represents the results of baseline models, dashed lines represents the results of FairSWAP. Higher and smoother lines indicate higher fairness. Models trained on separate gender groups possess more discrepancies than the same PAD models trained on fused data.

Metric	Trained	LBP [63]		ResNet50 [108]		DeepPixBis [94]		LMFD-PAD [83]	
		B	FairSwap	B	FairSwap	B	FairSwap	B	FairSwap
FDR	Fused	0.821	**0.870**	0.882	**0.900**	0.900	**0.904**	**0.902**	0.895
	Male	0.775	**0.820**	0.871	**0.895**	0.886	**0.888**	0.883	**0.889**
	Female	0.871	**0.901**	**0.893**	0.874	0.881	**0.885**	**0.889**	0.882
	Average	0.822	**0.864**	0.882	**0.890**	0.889	**0.892**	**0.891**	0.889
ABF	Fused	0.695	**0.762**	0.877	**0.905**	0.899	**0.909**	**0.901**	**0.901**
	Male	**0.443**	0.358	0.857	**0.901**	0.885	**0.890**	0.885	**0.899**
	Female	0.723	**0.894**	**0.891**	0.848	0.871	**0.888**	0.882	**0.883**
	Average	0.620	**0.671**	0.875	**0.885**	0.885	**0.896**	0.889	**0.894**

Table 3.3.: Fairness in terms of FDR-AUC and ABF-AUC with respect to gender. The higher FDR-AUC and ABF-AUC indicate a fairer PAD (highest in bold). FairSWAP enhances the fairness in most cases.

value 3.50% and 3.89%, while the lowest EER values obtained by models trained on females are 10.95% by DeepPixBis and 9.85% by LMFD-PAD. Both observations suggest that a model tends to learn a better feature representation from male group and shows a better PAD generalizability on male group, which is consistent with the observations in FR [150, 228].

Figure 3.6 explores the PAD performance under different ODTA s by illustrating the 1-BPCER(τ) values at various thresholds $\tau = APCER_x$ separately computed from fused, male, and female groups. By observing Figure 3.6, we conclude that 1) When evaluating models trained on fused data, using a decision threshold computed on female samples shows a relatively lower performance than a threshold from fused or male groups. 2) When evaluating models trained on males, the PAD performance is consistently higher (i.e, higher 1-BPCER values) on male test data than on females irrespective of ODTA s. 3) However, when evaluating models trained on females, male test data obtains comparable (by ResNet50 and DeepPixBis) and even higher PAD performance (by LBP-MLP) than testing on females, unlike the previous EER results observation that demographic data known in the training set exhibits better PAD performance. These observations indicate that in addition to the bias caused by the training data, the ODTA is also one of the triggers of PAD bias.

Fairness with respect to gender The above analysis focuses on the exploration of differential performance and outcome of PAD induced by training data and ODTA s. To further assess the training data-related fairness and performance in PAD, we illustrate ABFs in Figure 3.7. In addition, Table 3.3 presents the area under FDR (FDR-AUC) and area under ABF (ABF-AUC) of each PAD system. As suggested in [64], a more stable and higher FDR curve and a higher FDR-AUC refer to a fairer system. This section first focuses on the fairness of baseline models without the proposed FariSWAP, i.e. solid curves. We observe that all ABF curves in Figure 3.7 showed varied degrees of fluctuation. Moreover, deep-learning-based models trained on separate gender groups (blue and green curves) have more discrepancies than the same PAD model trained on fused data (red curves) in most cases. Such observations are consistent with the ABF-AUC values in Table 3.3. In addition, deep-learning-based models exhibit slightly higher FDR-AUC values than the handcrafted-feature-based approach. when associating PAD performance and fairness, deep-learning-based models obtain significantly higher ABF-AUC values in Table 3.3. To summarize, we concluded that the training data distribution does affect the gender fairness of face PAD methods.

3.6.2. Fairness assessment over occlusion (Protocol-2)

Similar to above gender fairness discussion, we report the face PAD performance in terms of EER values on each occlusion/non-occlusion group in Table 3.4 and the 1-BPCER(τ) value at different ODTA s in Figure 3.8. Furthermore, we assess the fairness of PAD models trained on various composition of training data in Figure 3.9 and Table 3.5.

PAD Performance with respect to occlusion As in Table 3.4, we observe that: 1) Training model on separate occlusion/non-occlusion data performs better on the learned group data. 2) Models trained on non-occlusion group obtain significant higher EER values than models trained on occlusion group on both test groups, respectively. 3) Models trained on fused data achieve slightly higher EER values on occlusion group in most cases.

Trained	Test	LBP [63]			ResNet50 [108]			DeepPixBis [94]			LMFD-PAD [83]		
		B	FairSwap	Impro	B	FairSwap	Impro	B	FairSwap	Impro	B	FairSwap	Impro
Fused	O	**13.40**	11.61	-1.79	**2.93**	**2.65**	-0.28	1.19	**1.46**	+0.27	**2.54**	**2.28**	-0.26
	w/o O	13.29	**12.4**	-0.89	2.04	1.82	-0.22	**1.33**	0.95	-0.38	1.98	1.67	-0.31
O	O	11.11	15.05	+3.94	3.64	2.68	-0.96	1.21	1.87	+0.66	2.87	2.80	-0.07
	w/o O	**14.71**	**18.3**	+3.59	**6.27**	**2.83**	-3.44	**2.69**	**3.43**	+0.74	**3.21**	**3.22**	+0.01
w/o O	O	**22.60**	**21.43**	-1.17	**11.40**	**7.89**	-3.51	**9.91**	**6.62**	-3.29	**8.94**	**8.19**	-0.75
	w/o O	19.46	19.60	+0.14	4.94	4.31	-0.63	6.08	3.86	-2.22	4.82	4.68	-0.14

Table 3.4.: PAD performance in terms of EER (%) on occlusion groups by baseline models (B) and FairSWAP. Fused, O, and w/o O refer to the fused data, occlusion, and non-occlusion data, respectively. Bold numbers indicate the highest EER values between occlusion and non-occlusion test data by each trained model. The results imply that occlusion group is more challenging to be correctly classified than non-occlusion group, but can help models learn more complex and generalized representations.

This might suggest that occluded data is harder to classified correctly than non-occluded data, but it can help models to learn a more complex and generalized representations. Furthermore, as in Figure 3.8, we empirically notice that the performance of models trained on fused data is decreased by assigning decision thresholds τ driven from occluded data. In addition, when τ at smaller APCER, LBP-MLP and DeepPixBis model perform worse on the non-occluded group than on the occluded group (i.e. curves of orange series lower than curves of blue series), which is in contrast to previous observations in terms of EER values. Besides, despite some performance degradation of models trained on separate group data by various thresholds, the tendencies of differential outcome between two test groups remain coincident with the observations in terms of EER values. Overall, we concluded that the training data, distribution different data and ODTA s are triggers of the PAD performance bias.

Fairness with respect to occlusion To neglect the effect of ODTA s, two fairness metrics (FDR and ABF) of PAD models on different training data groups are computed by applying a 'fair' decision threshold (as stated in the Section 3.4.2 and in [64]). Figure 3.9 illustrates the ABF values. In terms of the ABFs, it can be seen that deep-learning-based models trained on non-occlusion group are significant unfairer (solid green curves) than models trained on occlusion group (solid blue curves) and fused data group (solid red curves). With the fused training data, models show the smaller fairness discrepancies, i.e. higher and stable ABF curves. ABF-AUC values in Table 3.5 support these findings, i.e. highest ABF-AUC values are achieved by models trained on the fused data. ABF-AUC values are consistent with these observations and confirm that diverse training data is able to enhance

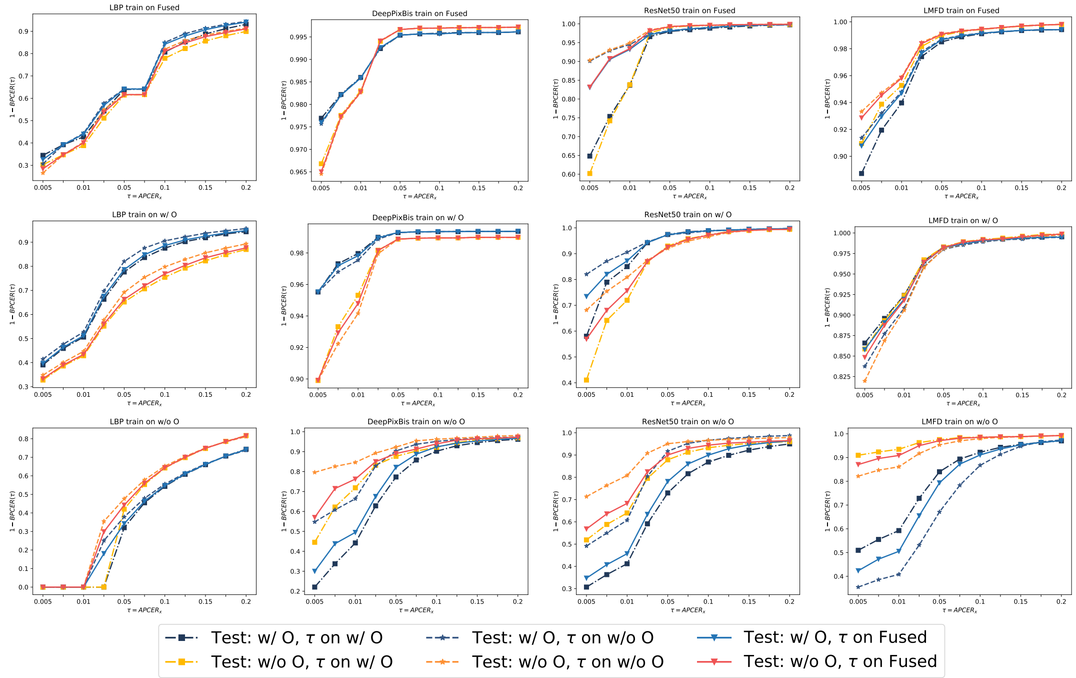


Figure 3.8.: The results of baseline models on occlusion group in terms of $BPCER@APCER_x$. To study the impact of ODTA on fairness, the threshold τ is calculated separately on fused, occlusion, and non-occlusion test set. Performance of models trained on fused data degrades when assigning a τ driven from the occlusion group.

Metric	Trained	LBP [63]		ResNet50 [108]		DeepPixBis [94]		LMFD-PAD [83]	
		B	FairSwap	B	FairSwap	B	FairSwap	B	FairSwap
FDR	Fused	0.887	**0.892**	0.898	**0.899**	0.904	**0.905**	0.901	**0.902**
	Occlusion	0.856	**0.884**	0.883	**0.899**	**0.901**	0.891	**0.901**	0.894
	w/o Occlusion	**0.875**	0.841	0.832	**0.834**	0.844	**0.865**	0.815	**0.849**
	Average	**0.873**	0.872	0.871	**0.877**	0.883	**0.887**	0.872	**0.882**
ABF	Fused	**0.874**	0.839	0.898	**0.904**	0.904	**0.908**	0.902	**0.907**
	Occlusion	0.828	**0.830**	0.874	**0.905**	**0.901**	0.897	0.900	**0.901**
	w/o occlusion	**0.826**	0.571	0.781	**0.794**	0.794	**0.861**	0.751	**0.826**
	Average	**0.843**	0.747	0.851	**0.868**	0.866	**0.889**	0.851	**0.878**

Table 3.5.: Fairness in terms of FDR-AUC and ABF-AUC with respect to occlusion attribute. The higher FDR-AUC and ABF-AUC in bold indicate a fairer PAD model.

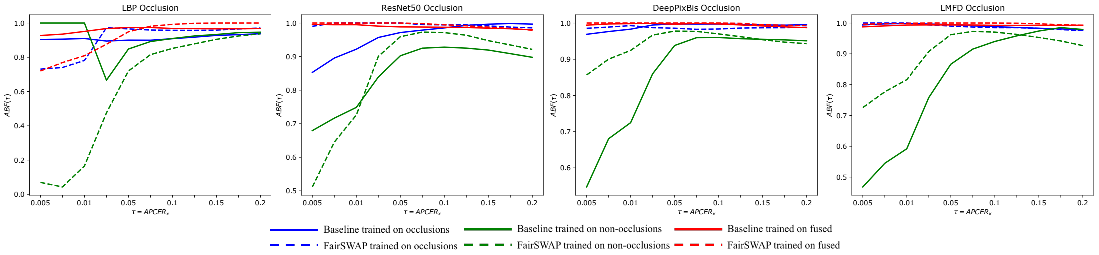


Figure 3.9.: Fairness measure in terms of ABF values by employing four face PAD models on occlusion groups. Solid lines represents the results of baseline models, dashed lines represents the results of FairSWAP. Higher and smoother lines indicate higher fairness. Models trained on fused and occluded data exhibit higher fairness.

both the performance and fairness of PAD algorithms.

Attribute	Type	LBP [63]			ResNet50 [108]			DeepPixBis [94]			LMFD-PAD [83]		
		B	FairSwap	Impro	B	FairSwap	Impro	B	FairSwap	Impro	B	FairSwap	Impro
Bangs	w	**13.42**	**12.70**	-0.72	**3.71**	**3.49**	-0.22	**1.36**	**1.60**	+0.24	**2.64**	**2.54**	-0.10
	w/o	13.19	12.05	-1.14	2.31	1.54	-0.77	1.27	0.92	-0.35	1.91	1.57	-0.34
Beard	w	12.16	8.21	-3.95	**2.96**	**2.08**	-0.88	0.42	0.46	+0.04	1.98	1.78	-0.20
	w/o	**13.33**	**12.39**	-0.94	2.65	2.05	-0.60	**1.38**	**1.16**	-0.22	**2.17**	**1.85**	-0.32
Eye glass	w	**13.61**	11.41	-2.20	**3.25**	**2.28**	-0.97	1.25	**1.61**	+0.36	**2.61**	**2.25**	-0.36
	w/o	13.45	**12.87**	-0.58	1.96	1.86	-0.10	**1.31**	0.67	-0.64	1.76	1.44	-0.32
Makeup	w	12.20	**14.12**	+1.92	**2.69**	1.34	-1.35	**2.49**	0.93	-1.56	**4.16**	1.19	-2.97
	w/o	**13.33**	11.95	-1.38	2.68	**2.12**	-0.56	1.22	**1.11**	-0.11	1.99	**1.76**	-0.23
LongHair	w	**21.69**	**16.80**	-4.89	2.50	**2.88**	+0.38	**1.82**	**1.84**	+0.02	**3.44**	**2.84**	-0.60
	w/o	11.11	11.10	-0.01	**2.67**	1.67	-1.00	1.19	1.05	-0.14	1.80	1.61	-0.19
StraightHair	w	12.90	12.09	-0.81	**2.82**	**2.17**	-0.65	**1.38**	**1.24**	-0.14	**2.30**	**1.92**	-0.38
	w/o	**14.36**	**12.81**	-1.55	1.68	1.19	-0.49	0.67	0.09	-0.58	0.83	0.97	+0.14
Average		13.73	12.38	-1.35	2.66	2.06	-0.60	1.31	1.06	-0.26	2.30	1.81	-0.49

Table 3.6.: PAD performance in terms of EER (%) on the other attribute groups by baseline (B) and FairSWAP models trained on fused data. Type w and w/o refer to test samples with such attribute, and without such attribute, respectively. Bold numbers indicate the highest EER values between w and w/o test attribute group by each trained model. Non-demographic attributes exhibit an implicit effect on PAD to some extent.

3.6.3. Fairness assessment over other attributes (Protocol-3)

In addition to the gender and occlusion attributes, we assess the PAD performance and fairness on six pairwise non-demograpihc attributes: with/without bangs, beard, eye-

Metric	Tested	LBP [63]		ResNet50 [108]		DeepPixBis [94]		LMFD-PAD [83]	
		B	FairSwap	B	FairSwap	B	FairSwap	B	FairSwap
FDR	Bangs	*0.893*	*0.895*	*0.900*	0.890	0.901	*0.901*	0.895	*0.899*
	Beard	0.832	0.834	*0.889*	*0.902*	0.890	0.898	*0.901*	0.897
	Eyeglasses	*0.883*	*0.899*	0.884	0.895	0.899	0.898	*0.902*	*0.902*
	Makeup	0.859	0.859	0.870	*0.901*	0.893	*0.901*	0.893	0.887
	LongHair	0.814	0.864	0.878	0.894	*0.902*	0.897	0.895	0.889
	StraightHair	0.850	0.882	0.877	0.889	*0.905*	*0.901*	0.885	0.896
	Average	0.855	**0.872**	0.883	**0.895**	0.898	**0.899**	**0.895**	**0.895**
ABF	Bangs	*0.884*	*0.837*	*0.900*	0.895	0.900	*0.906*	0.893	*0.905*
	Beard	0.779	0.647	*0.886*	*0.907*	0.887	0.904	*0.900*	0.902
	Eyeglasses	*0.872*	*0.879*	0.881	0.899	0.898	0.902	*0.902*	*0.907*
	Makeup	0.743	0.399∗	0.858	*0.906*	0.891	*0.906*	0.891	0.892
	LongHair	0.700	0.795	0.872	0.901	*0.901*	0.903	0.893	0.894
	StraightHair	0.803	0.816	0.873	0.894	*0.905*	0.894	0.893	0.902
	Average	**0.797**	0.729	0.878	**0.900**	0.897	**0.903**	0.895	**0.900**

Table 3.7.: Fairness in terms of FDR-AUC and ABF-AUC with respect to other attributes by employing models trained on fused data. The higher FDR-AUC and ABF-AUC in bold indicate a fairer PAD model. The bold numbers refer to the highest values between baseline and FairSWAP. FairSWAP enhances the fairness of all deep-learning-based PAD models.

glasses, makeup, long hair, and straight hair. We omit the assessment of the imbalanced training data due to the insufficient data across such attributes. The PAD performances on these attributes are evaluated only by adopting the models trained on the fused data.

PAD Performance with respect to attributes Table 3.6 presents the PAD performance in terms of the EER values (first focusing on the baseline results (B)). The bold numbers refer to the highest EER value between each paired attribute for each PAD approach. In case of the attribute *Bangs*, EER values of samples with bangs is consistently lower than without bangs for all training models. Subjects wearing eyeglasses are more difficult to detect correctly. For attribute *Beard*, a different behaviour is observed in comparison with other two occlusion annotations (*Bangs* and *Eyeglasses*), i.e. models perform better on samples with beards in most cases. Similarly, samples without makeup obtains lower EER values in most cases. Both may be attributed by that attributes without *Beard* and with *Makeup* correlate exclusively with the female group as presented in Figure 3.4. For attributes *LongHair* and *StaightHair*, samples with such attributes achieve higher error rates than samples without such attributes in most cases. By empirically observing the cropped faces, one possible reason is that long hair has slightly higher possibility to cover the ear region than short hair and thus leading to some discriminate patterns loss. Additionally, the pattern of curly hair may be somehow easier to be learned by the model than straight hair. Overall, these attributes are interrelated and entangled and the reasons behind the PAD behaviour are thus difficult to analyze in a stand-alone manner. We can conclude that non-demographic attributes also exhibit somehow implicit influences on PAD performance.

Fairness with respect to other attributes Table 3.7 presents FDR-AUC and ABF-AUC, where the italic numbers refer to the two highest FDR-AUC values and two highest ABF-AUC values. From Table 3.7, we make several observations: 1) Occlusion attribute (bangs, beard or eyeglasses) obtained the highest FDR-AUC and highest ABF-AUC values in most cases (six out of eight baseline cases). 2) Deep-learning-based PAD solutions are fairer than hand-crafted feature-based approach, especially when considering ABF values. This results may be explained by the fact that deep learning models learn more abstract features beyond the texture features. Overall, PAD models show the slightly different fairness between non-demographic attribute groups. These differences might be inherent from the training data that the numbers of training samples in each attribute groups are different.

3.7. Results of fairness enhancement

The fairness investigation results on CAAD-PAD (Section 3.6) showed that the training data and ODTA are triggers of PAD unfairness. The observed unfairness may be caused by PAD models learning the identity and semantic information beyond the discriminative attack cues. Therefore, we proposed the FairSWAP method, acting as a data augmentation technique, to enhance the PAD fairness (Section 3.5). This section discusses fairness enhancement results by applying the FairSWAP solution on gender, occlusion, and the remaining attribute groups of CAAD-PAD.

3.7.1. Fairness enhancement over gender (Protocol-1)

Table 3.2 presents the PAD performance in terms of EER values and Figure 3.10 illustrates the performance under different ODTAs. As shown in Table 3.2, three deep-learning-based PAD methods with the FairSWAP solution achieve improved performance in 13 out of 18 experimental cases on gender groups. However, LBP-MLP method did not benefit from the FairSWAP solution. The possible reason is that the mixed illuminations in augmented images decrease the discrimination ability of color-channel-based LBP features. In Figure 3.10, the PAD performance differences between male and female groups still exist, but the differences under different ODTAs are smaller, i.e. the curves of similar colors (blue series or orange series) are closer.

The dotted curves in Figure 3.7 refer to ABF curves and Table 3.3 presents the FDR-AUC and ABF-AUC values on gender groups. As shown in Figure 3.7, the dotted ABF curves are higher than solid curves in some cases (44 higher FDR values out of 72 cases by considering FDR at all plotted decision thresholds), especially when deep-learning-based models trained on the non-occlusion data (green curves). Although LBP-MLP with FairSWAP did not obtain improved PAD performance, it enhances the fairness LBP-MLP (i.e. higher FDR values) as shown in Table 3.2. PAD methods with FairSWAP achieve consistently higher ABF values than baseline models, indicating the combined benefits of FairSWAP in terms of performance and fairness.

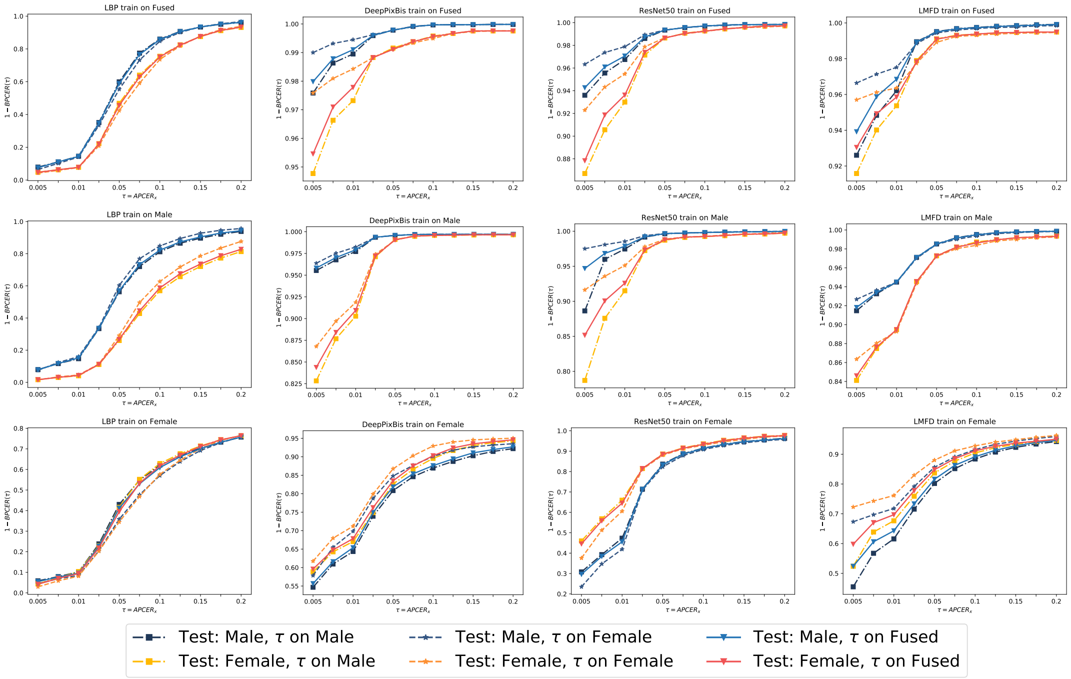


Figure 3.10.: The results of FairSWAP models on gender group in terms of $BPCER @ APCER_x$. While PAD performance differences between males and females still exist, the differences under different ODTAs are smaller, i.e. the curves within the blue and orange series are closer.

3.7.2. Fairness enhancement over occlusion (Protocol-2)

As shown in Table 3.4 and Figure 3.11, we conclude that: 1) LBP-MLP method does not benefit much from the FairSWAP solution compared to the deep-learning-based models.

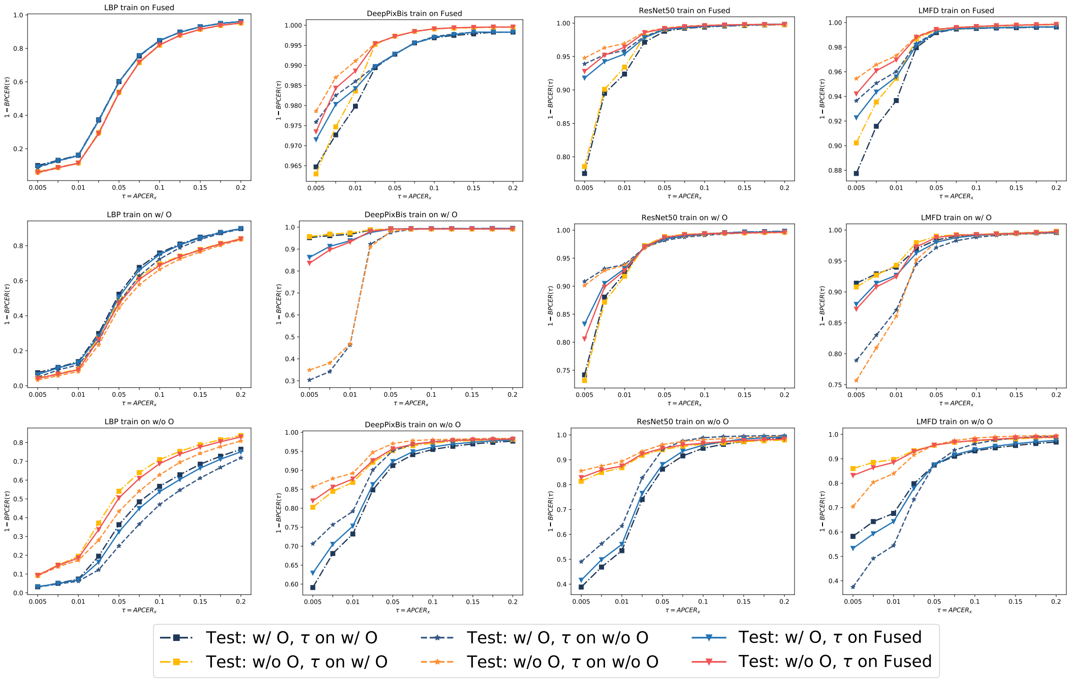


Figure 3.11.: The results of FairSWAP on occlusion group in terms of $BPCER$ @ $APCER_x$. With the help of FairSWAP, the difference in PAD performance under different ODTAs is reduced, i.e. the curves are closer within the blue and orange series.

2) When evaluating models trained on the non-occlusion group, models with FairSWAP significantly outperform baseline models (improvements in seven out of eight cases). 3) The differential performance and outcome of PAD under different ODTAs are decreased with the help of the FairSWAP solution. As shown in Figure 3.9, the dotted FDR curves are smoother than solid curves when LBP-MLP, DeepPixBis, and ReseNet50 are trained on the occlusion group (blue curves). In the case of models trained on the non-occlusion group, FDR curves of DeepPixBis and LMFD-PAD with FairSWAP (dotted green lines) are higher than baseline models (solid green lines). This visual observations are reflected in the ABF-AUC values in Table 3.5. Deep-learning-based methods with FairSWAP solution result in higher FDR-AUC values in all cases, and higher ABF-AUC values except LBP-MLP method achieves comparable results (0.873 FDR-AUC without FairSWAP and 0.872 with FairSWAP). In general, FairSWAP approach enhances the PAD performance and fairness on occlusion attribute in most cases, especially for deep-learning-based PAD solutions.

3.7.3. Fairness enhancement over other attributes (Protocol-3)

Table 3.6 compares the PAD performance of baseline models and models with FairSWAP in terms of EER. FairSWAP improves the PAD performance on most protocol-3 attribute groups. By considering the average performance of each model, models with FairSWAP outperform baseline models. The FDR-AUC and ABF-AUC values are presented in Table 3.7. FairSWAP obtains comparable or higher fairness (i.e. higher FDR-AUC values) than baselines. Looking at ABF-AUC values, deep-learning-based models benefit from the FairSWAP approach, while LBP-MLP with FairSWAP obtains a lower average ABF-AUC value than without FairSWAP, mainly due to the behavior with the Makeup attribute. Overall, the FairSWAP approach enhances the PAD performance and fairness on most attribute groups, especially for deep-learning-based PAD solutions.

3.8. Discussion

To enable answer *RQ1* and *RQ2*, this chapter presented CAAD-PAD dataset by combining several well-known PAD datasets where seven human-annotated attribute labels are provided. Then, to represent both the PAD fairness and the absolute PAD performance simultaneously, a novel metric, ABF, is introduced. As a result, the answers to *RQ1* and *RQ2* are provided with detailed investigations and analyses.

- This chapter conducted a set of extensive experiments following three designed experimental protocols over gender, occlusion, and the other attributes. With such experiments for fairness assessment, it can be concluded that training data and deployment choices, ODTA on different groups, are triggers of the unfairness of face PAD. For example, when models are trained separately on male and female groups, EER values are higher for the gender group that is unseen during the training phase. Moreover, models trained on fused data achieve consistently higher EER values on females than males. When evaluating models trained on fused data, using a decision threshold computed on female samples shows relatively lower performance than a threshold from fused or male groups.

- Based on these analyses, this chapter proposed a data augmentation method, FairSWAP, to mitigate the bias in the trained PAD solutions. Detailed experimental results demonstrate that FairSWAP approach enhanced the PAD performance and fairness on gender, occlusion, and other attributes in most cases, especially for deep-learning-based PAD solutions.

3.9. Summary

To address the understudied issue of fairness in face PAD, this chapter provided a comprehensive analysis of the topic. To enable the answering of *RQ1* and *RQ2*, the chapter first presented the CAAD-PAD dataset, which combines six publicly available face PAD datasets in Section 3.3. The dataset includes seven human-annotated attribute labels covering both demographic and non-demographic attributes. Additionally, the chapter introduced a novel metric, ABF, which jointly represents both the PAD fairness and absolute PAD performance. This chapter analyzed the fairness of face PAD and its relation to the nature of the training data and the ODTA on different data groups by employing four face PAD methods. Extensive experimental results in Section 3.6 pointed out that the training data and ODTA are triggers of unfairness in face PAD, providing an answer to *RQ1*. Consequently, this chapter proposed a simple yet effective solution, FairSWAP, acting as a data augmentation technique, to enhance the fairness of face PAD. FairSWAP aims to disorder the attribute information and guide models to mine discriminative attack features instead of identity or appearance features. Extensive experiments on the CAAD-PAD dataset in Section 3.7 demonstrated that FairSWAP boosts PAD performance and fairness on gender, occlusion, and other attribute groups in most experimental setups, providing an answer to *RQ2*.

The next chapter focuses on the emerging challenge of masked face PAD.

4. The emerging challenge of masked face presentation attack detection

The previous chapter assessed the fairness of the face PAD algorithms and alleviated the bias by intelligently augmenting the training data. This chapter addresses the emerging and unexpected challenge for FR and PAD posed by wearing a facial masked during the COVID-19 pandemic. This chapter will first present a new large-scale face masked PAD database to enable the study of the effect of facial masked PAs on FR and PAD systems. Then, this chapter provides a comprehensive analysis of the effect of masked face PAs for the existing FR and PAD systems as a response to *RQ3* and *RQ4*. Furthermore, this chapter introduces a partial attack label supervision and regional weighted inference-based method to address the face masked PAs issue as a response to *RQ5*. This chapter is based on the published papers [75, 84].

4.1. Introduction

Since the SARS-CoV-2 coronavirus outbreak and its rapid global spread, wearing a mask has become one of the most efficient ways to protect and prevent getting infected with COVID-19. However, for identity checks in crowded scenarios such as at airports, removing the mask for FR increases the chance of infection. Wearing masks in public might be an essential health measure and a new norm even after the COVID-19 pandemic as most countries support the use of masks to minimize the spread of the virus. As a result, researchers have shown an increased interest in the effect of face masks on the performance of FR verification [51, 54, 181]. The results of their studies have shown that pre-COVID-19 FR algorithms suffer performance degradation owing to the masked faces. However, attacks compromising the security and vulnerability of FR systems for subjects wearing face masks have so far been overlooked. In this study, security refer to the presentation attacks (PAs). Attackers can use PAs to spoof FR systems by impersonating someone or obfuscating their identity. Common PAs include printed photos/images, replayed videos and 3D masks [134, 136]. Driven by the ongoing COVID-19 pandemic, presentation attack detection (PAD) [250] has encountered several understudied challenges when facing masked faces. Current face PAD databases [14, 168, 169] only contain printed images or replayed videos in which subjects were not wearing face masks. Therefore, there is uncertainty about the relationship between the performance of PAD techniques and PAs with face masks. Moreover, the vulnerability of FR systems to masked attacks remains unclear. To overcome such gaps, researchers require well-studied masked PAs. In this study, we design and collect three types of attacks based on masked and unmasked face images collected realistically and collaboratively [51, 54]. The bona fide samples were divided into categories of BM0 (subjects wearing no masks) and BM1 (subjects wearing masks). AM0 data are unmasked print/replay attacks, which are commonly used data in most current PAD databases. AM1 data include print/replay attacks, where live subjects wore masks. In addition, we provide a novel partial attack type, called AM2, where a real medical mask is placed on printed photos or replayed videos to simulate the subject wearing a mask. This is motivated by our assumption that AM2 might be a challenging attack as it contains both bona fide and attack presentations that may confuse PAD and/or FR systems. The data samples are presented in Figure 4.1.

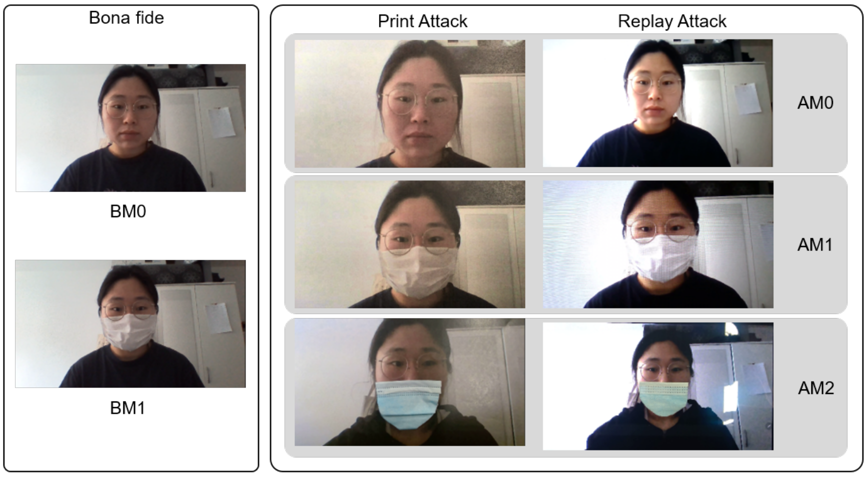


Figure 4.1.: Example bona fide and attack samples in the CRMA database. Based on the presence of face masks, bona fides are grouped into BM0 (without mask) and BM1 (with mask) categories. The novel attacks are grouped into AM0 (spoof face without mask), AM1 (spoof face with mask), AM2 (spoof faces covered by real masks).

To address the above concerns, this chapter introduced a new large-scale face PAD database, investigated and analyzed the effect of masked face attacks on FR and PAD methods. Additionally, a PAD solution was proposed to enhance the PAD performance on masked face attacks. In the efforts seeking to answer *RQ3*, *RQ4*, and *RQ5*, the main contributions in this chapter are:

- The novel CRMA was presented in this chapter. Three types of PAs, called AM0 (unmasked face PA), AM1 (masked face PA), and AM2 (unmasked face PA with a real masked placed on the PA) (as shown in Figure 4.1), was created for both print and replay PAIs. To create such attacks, three electronic tablets with high-resolution and three capture scales are used. Additionally, we designed three experimental protocols to explore the effect of masked attacks on PAD performance.

- As a response to *RQ3*, in-depth vulnerability analysis of FR systems was presented. Three deep-learning-based FR techniques for three types of PAs were evaluated. The experimental results indicated that these three FR networks exhibit significantly higher vulnerabilities to the real mask attacks than masked face attacks.

- As a response to *RQ4*, extensive experiments were conducted to explore the effect of bona fide samples, masked faces attacks, and real masks (on spoof faces) on the face PAD behavior. To support a comprehensive evaluation, seven PAD algorithms, comprising texture-based, deep-learning-based, and hybrid methods, were selected to evaluate the PAD generalizability in intra- and cross-database scenarios under three mask-related protocols. Both quantitative and qualitative analyses revealed that masked bona fides and PAs dramatically decreased the performance of PAD algorithms. Moreover, deep-learning-based methods perform worse on real mask attacks than mask-face attacks in most cases.

- As a response to *RQ5*, a PAD method was proposed that considers partial attack labels to supervise the PAD model training, as well as regional weighted inference, to further improve the PAD performance by varying the focus on different facial areas. The proposed method was not directly linked to specific network architecture and thus can be directly incorporated into any common or custom-designed network. The proposed method outperformed seven established PAD methods on the CRMA database by reducing the mentioned shortcomings when facing masked faces. Additionally, a detailed step-wise ablation study pointed out the individual and joint benefits of the proposed concepts on the overall PAD performance.

In the rest of the chapter, a brief review of relevant works is provided in Section 4.2. Then, the new CRMA database is described in detail in Section 4.3. The proposed PAD solution is introduced in Section 4.4. Section 4.5 presents the experimental setups and implementation details. Section 4.6 presents and discusses the achieved results. Finally, a set of conclusions are drawn in Section 4.7.

4.2. Related work

This section reviews the most relevant prior works from three perspectives: face PAD databases, face PAD methods, and FR and vulnerability analysis. At the end of each part, the difference between our work and prior work is pointed out.

Face PAD databases: Data resources have become especially important ever since the advent of deep learning, because machine-learning-based algorithms have the risk of underfitting or overfitting on limited data. Given the significance of good-quality databases, several face PAD databases have been released, such as NUAA [214], CASIA-FAS [261], Replay-Attack [43], MSU-MFSD [233], OULU-NPU [14], and SiW [168], all consisting of 2D print/replay attacks. In addition, SiW-M [169] and CelebA-Spoof [259] databases provide multiple types of attacks such as makeup, 3D mask, or paper cut. Moreover, some multimodal databases are publicly available: 3DMAD [71], Mssproof [45], CASIA-SURF [258], and CSMAD [9]. These databases undoubtedly contribute to the significant progress of PAD research. For example, the CeleA-Spoof database comprises images from various environments and illuminations with rich annotations to reflect real scenes. However, these databases also have weaknesses: 1) The multi-modal databases have high hardware requirements and cannot be widely used in daily life; 2) Some databases such as CASIA-MFS [261] and MSU-MFS [233] cannot satisfy the current needs because of the lower quality of the outdated acquisition sensors; 3) Oulu-NPU [14], SiW [168], SiW-M [169], and CelebA-Spoof [259] are relatively up-to-date, but they do not consider PAs with real face masks to fit the current COVID-19 pandemic. Hence, we collect the CRMA database to fill the gaps in these databases in the context of the ongoing COVID-19 pandemic; Furthermore, we ensure the database is generalizable and compatible with real scenarios. The CRMA database can be used to better analyze the effect of a real mask on PAD performance and the vulnerability of FR systems for novel attacks, such as placing a real mask on an attack presentation. Detailed information related to the databases mentioned above is presented in Table 4.1).

Face PAD methods: In recent years, there has been an increasing number of studies in the field of face PAD [87, 140, 210]. These studies can be broadly grouped into three categories: texture-based methods, deep-learning-based methods, and hybrid methods. Texture features, such as LBP [183], project the faces to a low-dimensional embeddings. Määttä *et al.* [173] proposed an approach using multi-scale LBP to encode the micro-texture patterns into an enhanced feature histogram for face PAD. The resulting histograms were

Database	Year	# Subjects	# Data (BF/attack)	Capture devices (BF/attack)	Display devices	Modality	Attack type
NUAA [214]	2010	15	5105/7509 (I)	Webcame	-	RGB	1 Print
CASIA-FAS [261]	2012	50	150/450 (V)	Two USB cameras, Sony NEX-5	iPad	RGB	1 Print, 1 Replay
Replay-Attack [43]	2012	50	200/100 (V)	MacBook 13 / iPhone 3GS, Cannon SX150	iPhone 3GS, iPad	RGB	1 Print, 2 Replay
3DMAD [71]	2013	17	170/85 (V)	Microsoft Kinect	-	RGB/Depth	1 3D Mask
Msspoof [45]	2015	21	1,680/3,024 (I)	uEye camera	-	RGB/IR	1 Print
MSU-MFSD [233]	2015	35	110/330 (V)	MacBook Air, Google Nexus 5 / Cannon 550D, iPhone 5s	iPad Air, iPhone 5s	RGB	1 Print, 2 Replay
Oulu-NPU [14]	2017	55	1,980/3,960 (V)	6 smartphones	Dell 1905FP, Macbook Retina	RGB	2 Print, 2 Replay
SiW [168]	2018	165	1,320/3,300 (V)	Cannon EOS T6, Logitech C920 webcam	iPad Pro, iPhone 7, Galaxy S8, Asus MB 168 B	RGB	2 Print, 4 Replay
CASIA-SURF [258]	2018	1000	18000/3000 (I)	RealSense camera	-	RGB/IR/Depth	5 Papercut
CSMAD [9]	2018	14	88/160 (V)	RealSense, Compact Pro, Nikon P520	-	RGB/IR/Depth/LWIR	1 silicone mask
SiW-M [169]	2019	493	660/1630 (V)	Logitech C920, Cannon EOS T6	-	RGB	1 Print, 1 Replay, 5 3D mask, 3 Makeup, 3 Partial
Celeb-Spoof [259]	2020	10,177	202,559/475,408 (I)	Various cameras/ 20 smartphones, 2 webcams, 2 tablets	PC, phones, tablets,	RGB	3 Print, 3 Replay, 1 3D mask, 3 Paper Cut
CRMA	2021	47	423/12,690 (V)	Webcams/iPad Pro, Galaxy Tab S6, Surface Pro 6	iPad Pro, Galaxy Tab S6, Surface Pro 6	RGB	1 Print, 3 Replay, **1 Real mask**

Table 4.1.: The summary of face PAD databases, including our CRMA database information for brief comparison. It should be noted that our CRMA database is the only database containing subjects wearing face masks and real face mask attacks. The details of our CRMA database are presented in Section 4.3.

then fed to a Support Vector Machine (SVM) classifier to determine whether a sample is a bona fide or attack. The LBP features extracted from different color spaces [16] were further proposed to utilize chrominance information. They achieved competitive results on Replay-Attack [43] (EER value of 0.4%) and CASIA-FAS [261] (EER value of 6.2%) databases. Furthermore, Boulkenafet *et al.* [13] organized a face PAD competition based on the OULU-NPU database and compared 13 algorithms provided by participating teams and one color-LBP-based method (referred to as baseline in [13]). In this competition, the GRADIANT algorithm fused multiple information, that is, color, texture, and motion. The GRADIANT achieved competitive results in the four evaluation protocols. In addition to the texture-based GRADIANT approach, deep-learning-based method (MixFASNet) or hybrid method (CPqD) also achieved lower error rates in all experimental protocols. CPqD fused the results from the fine-tuned Inception-v3 network and the color-LBP-based method (referred to as the baseline in [13]). Consequently, we chose to re-implement the color-LBP and CPqD methods in this study (details in Section 4.5.4), while the GRADIANT and MixedFASNet are discarded in our work because they do not provide sufficient details for re-implementation. Deep-learning-based methods have been pushing the frontier of face PAD research and have shown remarkable improvements in PAD performance. Lucena *et al.* [172] presented an approach called FASNet in which a pre-trained VGG16 is fine-tuned by replacing the last fully connected layer. The FASNet network achieved excellent performance on 3DMAD [71] and Replay-Attack databases [43]. Recently, George *et al.* [93] proposed training a network with pixel-wise binary supervision on feature maps to exploit information from different patches. DeepPixBis [93] outperformed the state-of-the-art algorithms in Protocol-1 of the OULU-NPU database (0.42% ACER) but also achieved significantly better results than traditional texture-based approaches in the cross-database scenario. Considering the popularity of PAD techniques and the ease of implementation, we also chose FASNet and DeepPixBis (details in Section 4.5.4) to study the effect of the real mask and masked face attacks on the PAD performance. However, the performance degradation might be caused by the noisy pixel-wise label used in [93] when handling partial attacks. Kantarci *et al.* [144] improved the performance of the DeepPixBis method by shuffling the patches from input images and combining face patches. The generated new inputs improved the generalizability of the trained model under cross-database scenarios. Yu *et al.* [247] proposed a pyramid pixel-wise supervision method, which decomposed the pixel-wise label into multiple spatial scales for the supervision of multi-scale deep features. The pyramid supervision is able to interpret a richer spatial context, which is beneficial for fine-grained feature learning. Overall, pixel-wise labels have proven to be helpful in the improvement of PAD performance. Nevertheless, the quality of the pixel-wise labels is essential for the convergence of the trained networks. The coarse all zero or all one pixel-wise binary mask might not be suitable for the partial attack (masked face PAs in this case). To overcome the issues on partially masked attacks and masked face PAD in general, a PAL-RW solution is proposed by considering partial pixel-wise labeling and varying consideration of different facial areas (details in 4.4).

Face recognition and vulnerability analysis: As one of the most popular modalities, the face has received increasing attention in authentication/security processes, such as smartphone face unlocking and Automated Border Control (ABC). Moreover, FR techniques [28, 68, 167] have achieved significant performance improvements, and many personal electronic products have deployed FR technology. However, the ongoing COVID-19 pandemic brings a new challenge related to the behavior of collaborative recognition techniques when dealing with masked faces. Collaborative data collection refers to a subject actively attending to use the FR systems, such as unlocking personal devices or using an ABC gate, in contrast to uncollaborative capture scenario where the user does not intentionally use the FR service, such as in the case of surveillance. The NIST [181] provided a preliminary study that evaluated the performance of 89 commercial FR algorithms developed before the COVID-19 pandemic. Their results indicated that digitally applied face masks with photos decreased the recognition accuracy; for example, even the best of the 89 algorithms had error rates between 5% and 50%. It is worth noting that the masks used in the experiments were synthetically created. Damer *et al.* [51, 54] presented a real mask database to simulate a realistically variant collaborative face capture scenario. Each participant was asked to simulate a login scenario by actively looking toward a capture device, such as a static webcam or a mobile phone. Our attack samples were created and collected based on the masked face data, which refers to the bona fide samples in the PAD case (as described in Section 4.3). They also explored the effect of wearing a mask on FR performance and concluded that face masks significantly reduce the accuracy of algorithms. Mohammadi *et al.* [179] provided empirical evidence to support the claim that the CNN-based FR methods are extremely vulnerable to 2D PAs. Subsequently, Bhattacharjee *et al.* [9] presented the first FR-vulnerability study on 3D PAs. The experiments also clearly showed that CNN-based FR methods are vulnerable to custom 3D mask PAs. However, the vulnerability of FR systems to PAs with face masks has not been investigated. Therefore, in this chapter, we selected three CNN-based FR algorithms for further FR-vulnerability analysis on masked face attacks: the state-of-the-art ArcFace [68], SphereFace [167], and VGGFace [28]. These algorithms are discussed in more detail in Section 4.5.1.

4.3. The collaborative real mask attack database (CRMA)

The proposed CRMA database can serve as a supplement to the databases in Table 4.1, and because of the COVID-19 pandemic, it can better reflect the possible issues facing real-world PAD performance. The CRMA database includes 1) both unmasked (BM0) and masked (BM1) bona fide samples collected in a realistic scenario [51, 54], 2) conventional replay and print PAs created from faces not wearing a mask (AM0), 3) replay and printed PAs created from masked face images (AM1), and 4) novel PAs where the PAs of unmasked faces are covered (partially) with real masks (AM2), as shown in Figure 4.1. Damer *et al.* [51, 54] collected data to investigate the effect of wearing a mask on face verification performance. For PAD, such data are considered bona fide. The data presented in this study build on an extended version of the data introduced in [51, 54], by creating and capturing different types of PAs based on the bona fide data captured in [51, 54]. As a result, the bona fide data in this work are an extended version of the one introduced in [51, 54] and the attack data presented here are completely novel and have not been previously studied.

Figure 4.3 introduces the general statistical information of the CRMA database. This database contains 62% males and 38% females. The attack AM0, AM1, and AM2 ratios are 30%, 60%, and 10%, respectively, as will be described later in this section. Additionally, we count the frequency of the proportion of the face size in the video. The histogram shows that the proportion of the face areas in the videos is mostly between 5% and 30%. This section first describes the bona fide samples provided by [51, 54], and then introduces our process of attack sample creation and collection.

4.3.1. Collection of bona fide samples

To explore the FR performance on masked faces, Damer *et al.* [51, 54] recently presented a database where the subjects wearing face masks. This database simulates a collaborative environment in which participants collect videos by actively looking towards the capture device. During this process, the eyeglasses were removed when the frame was considered very thick following the International Civil Aviation Organization (ICAO) standard [125]. The videos were captured by the participants at their residences while working from home. Therefore, the types of face masks, capture devices, illumination, and background were varied. For PAD, these videos are classified as bona fides and will be used later to create attack samples.

The final version [51] of this database contains 47 participants. Each subject recorded a total of nine videos over three days with three different scenarios for each day. In contrast to the study by Damer *et al.* [54], which examined the effects of both face masks and illumination variations, we focused only on the impact of face masks on PAD performance. Hence, in our study, the bona fide videos are divided into two categories: a face without a mask on is denoted as BM0 (three videos per subject), and a face with a mask is marked as BM1 (six videos per subject) (as shown in the right column of Figure 4.1).

4.3.2. Creation of the presentation attacks

Most FR databases tried to collect data under various harsh conditions, such as poor lighting, strong occlusion, or low resolution. Such databases attempted to reproduce what might happen in a real-world scenario when a legitimate user obtains authorization [118]. In contrast, attackers use highly sophisticated artifacts, such as high-resolution images or videos, to maximize the success rate when impersonating someone. For this reason, we first collect the PAs in a windowless room where all lights are on. Second, three high-resolution electronic tablets were used in the acquisition process: 1) iPad Pro (10.5-inch) with the display resolution of 2224×1668 pixels, 2) Samsung Galaxy Tab S6 with the display resolution of 2560×1600 pixels, 3) Microsoft Surface Pro 6 with the display resolution of 2736×1824 pixels. In the process of collecting data, the capture devices and displayed images/tablets were stationary. The videos were captured with a resolution of 1920×1080. In addition, each video had a minimum length of 5 seconds, and the frame rate was 30 fps. This work focuses on the two common PAIs, print and replay attacks, due to their ease of creation and low cost. The attack data in each PAI (see the samples in Figure 4.1) are divided into three types: 1) the spoof face with no face mask (AM0), 2) the spoof face with a mask on (AM1), and 3) the spoof face with no face mask, but a real mask was placed on it to simulate a participant wearing a mask

(AM2). However, the size of the face area in each video is slightly inconsistent because the videos were recorded by the participants themselves. To reproduce the appearance of wearing a mask in the real world, we cropped five face masks to fit most of the faces (see Figure 4.3). The five masks are three small blue surgical masks, one slighter bigger white face mask, and one uncropped mask. When placing the mask, we select a suitable mask according to the size of the face in the printed image or video, aiming to cover the nose to the chin area and the cheeks without exceeding. The details of each PAI are as followings:
Print image attack: In print PAI, an attacker tries to fool the FR system using a printed photo. Considering the instability of the face during the first second, such as the participant pressing the recording button or adjusting the sitting position, the 35th frame of each bona fide video was printed out as an attack artifact. Therefore, we obtained nine photos per subject. The three tablets mentioned above were used to capture the photos. Furthermore, to increase the diversity and variety of the data, each tablet captured three videos for a photo with three scales (see examples in Figure 4.2). The captured videos using the first scale contained all areas (100%) of the photos, the second scale consisted of most areas (80%) of the original photos, and the third scale focused on the face area (60%) as much as possible. In addition to collecting attack data solely from printed images, we also collected data from real face masks overlaid on photos (i.e., the previously defined AM2). Theoretically, real masks will reduce the region of artificial features and increase the complexity and mixture of the features in the collected attack data. Eventually, 90 print attack videos were generated for each subject, that is, a total of 4,230 videos for 47 subjects in print PAI.
Replay video attack: In replay PAI, an attacker tries to obtain the authentication by replaying a video. The three common points of the collection process between print and replay PAI are the use of three tablets, the use of three scales, and the process of AM2 data creation, respectively. The difference is that these tablets were also used for capturing displays of videos (see examples in Figure4.2). While one tablet was replaying the video, the other two tablets were used to capture the data. As a result, each subject corresponded to 180 replay attack videos (162 videos of the AM0 and AM1 groups, 18 videos of AM2.), i.e., there were a total of 8,460 videos in this attack subset.

4.4. Methodology

This section describes the proposed method (PAL-RW) that utilizes our partial attack label supervision (PAL) and regional weighted inference (RW) for masked face PAD, along with the used backbone network architectures. The PAL intends to provide more accurate ground truth for partial attacks and thus enhances the convergence of the model training.

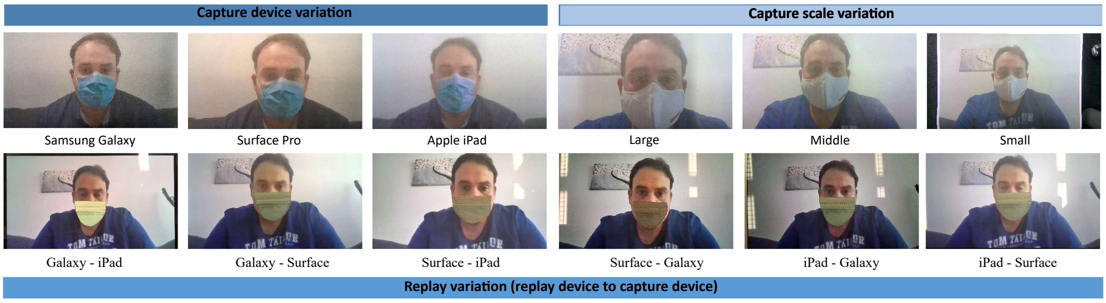


Figure 4.2.: Different capture variations in the CRMA database. The top left shows the videos captured by different devices. The top right shows the different capture scales. The bottom shows the six cross-device types of replay attack settings.

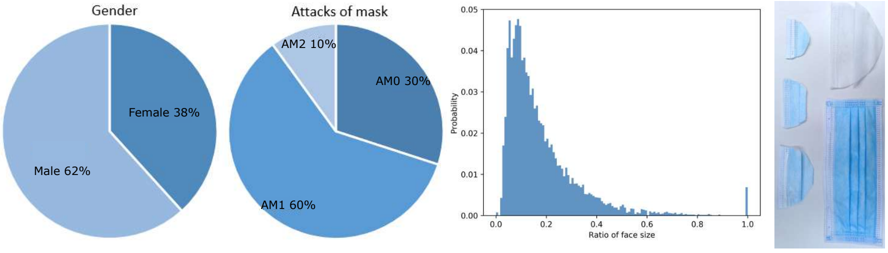


Figure 4.3.: The statistics of the subjects and the used mask shapes for creating AM2 samples in the CRMA database. From left to right: gender, mask types of attacks (AM0, AM1, AM2), the histogram shows the probability distribution of the face size ratio and the applied mask shapes.

Moreover, the RW post-processes the prediction results by increasing the focus on certain facial areas to further improve the PAD performance. Figure 4.4 depicts the training and testing phase in detail.

4.4.1. Partial attack label

Recent PAD works can be grouped into two classes: 1) global binary supervision where a network is supervised by a binary scalar label while training [13, 172] (as shown in Figure 4.5a), and 2) pixel-wise binary supervision where a network outputs a feature map and is supervised by binary mask, such as [65, 93, 144, 247]. Each pixel in the

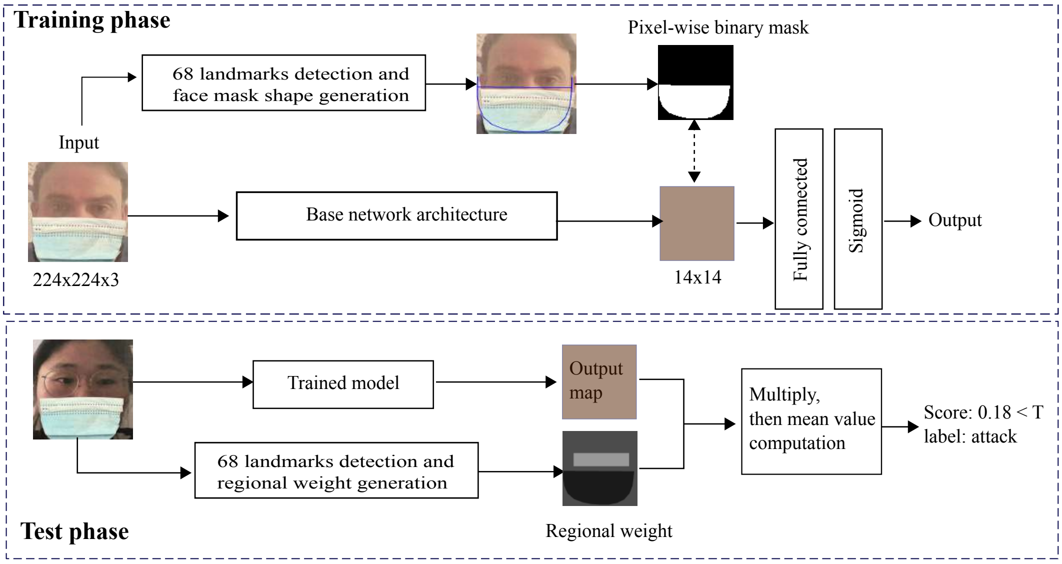
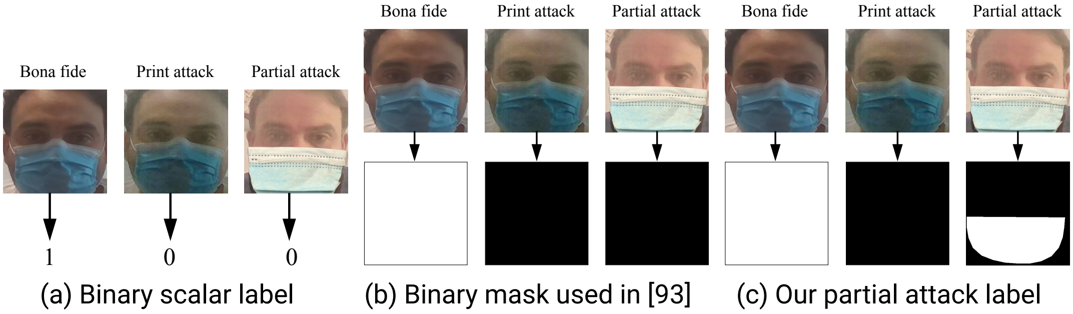

Training phase
Input
68 landmarks detection and face mask shape generation
Pixel-wise binary mask
Base network architecture
224x224x3
14x14
Fully connected
Sigmoid
Output
Trained model
Output map
68 landmarks detection and regional weight generation
Regional weight
Multiply, then mean value computation
Score: 0.18 < T
label: attack
Test phase

Figure 4.4.: An overview of our proposed PAL-RW method. The input of the model is a face image of $224 \times 224 \times 3$ pixels, and the model is supervised by a feature map with the size of 14×14 in addition to a binary output. In the training phase, the pixel-wise label for partial attack (AM2) is produced based on 68 facial landmarks. The pixel-wise label of the real face mask in AM2 data is set to 1 (bona fide) instead of a zero map for this attack. In the inference phase, the final PAD decision score is the mean value of the regional weighted feature map for further performance improvement. The lighter the color in the region weight map, the higher the weight value, that is, the eye region contributes more to PAD decision.

feature map is assumed as either bona fide (1) or attack (0). Most pixel-wise supervision methods [93, 144] utilized a zero map as ground truth of attacks and a one map for bona fide samples as shown in Figure 4.5b). They improved the performance on print and replay attacks, but performance degradation might occur when dealing with partial attacks. A possible reason is the incorrect pixel-wise labels of attacks. For example, the CRMA database contains an attack (AM2 in Figure 4.1) where a real mask was placed on unmasked spoof faces. To address this issue, we intend to make the pixel-wise label more accurate and thus enhance the training by partial attack label supervision. As shown in Figure 4.5c, the shape of the face mask is detected based on 68 facial landmarks, and each pixel-wise label in the face mask region is set to one (bona fide). Note that this operation on the binary mask is only performed for the partially masked face attack (AM2).



(a) Binary scalar label (b) Binary mask used in [93] (c) Our partial attack label

Figure 4.5.: Examples of three supervision methods. (a) Binary scalar labels for global supervision. For example, 1 represents bona fide and 0 represents attack [172, 173]. (b) Pixel-wise binary mask for local supervision such as used in [93]. A zero map is usually treated as ground truth of attack and a one map is for bona fide. (c) Our proposed partial attack label for AM2 attacks. The real face mask placed on an attack face is segmented and annotated as bona fide, while the rest of the attack is labeled as an attack.

4.4.2. Model architecture

To train a PAD solution based on our partial attack labels, we select two network architectures (DeepPixBis [93] and MixFaceNet [18]) as our backbones to validate our proposed PAL and RW on multiple architectures. MixFaceNet [18] is chosen as the second utilized backbone architecture because it is an extremely efficient architecture for face verification and identification, which possesses lower computation complexity (FLOPs) and high accuracy. MixFaceNet additionally contains different sizes of convolutional kernels, which might be beneficial to capture different levels of attack clues. The MixFaceNet is used with the weights initialized as the publically available weights pre-trained on the MS1MV2 dataset for face recognition, which might help maintain more subtle facial features. On the other hand, the DeepPixBis [93] architecture is used with the initial weights set by pre-training on the ImageNet dataset [67] for the general computer vision tasks, as described in [93]. DeepPixBis [93] method uses DenseNet [117] as a base network architecture and outputs a feature map of 14×14 pixels and a binary scalar prediction. The training of DeepPixBis is supervised by a binary mask and a binary label. MixFaceNet was partially inspired by the MixNets, and the channel shuffle operation to the MixConv block was introduced for enhancing the FR performance. In this case, the input image size of $112 \times 112 \times 3$ is changed to $224 \times 224 \times 3$ for outputting a feature map of 14×14 pixels, to be identical to the DeepPixBis backbone. Furthermore, the embedding stage of MixFaceNet is removed and replaced by two fully connected layers. The first fully connected layer is employed to output a feature map for pixel-wise supervision, while the second fully connected layer is used for binary classification. The MixFaceNet is also supervised by a binary mask and a binary label. Both models are trained by BCE loss function.

$$\mathcal{L}_{BCE} = -[y \cdot \log p + (1-y) \cdot \log(1-p)], \tag{4.1}$$

where y is the ground truth (1 for bona fide and 0 for attack in our case) and p is predicted probability. The overall loss equation for training of both models is shown below:

$$\mathcal{L}_{overall} = \lambda \cdot \mathcal{L}_{BCE}^{pixel-wise} + (1-\lambda) \cdot \mathcal{L}_{BCE}^{binary}, \tag{4.2}$$

where λ is set to 0.5 in our experiments.

4.4.3. Regional Weighted Inference

Once the model is trained, the probability of a given image can be estimated if it is an attack or bona fide. Most pixel-wise supervision-based methods use the mean value of the output feature map as the final decision score [79, 93], which neglects the differences in features at various facial regions. Therefore, we propose the regional weighted inference to post-process the prediction scores of models. The RW is motivated by two main previous observations: 1) the eye regions (including eyebrows) had a more significant influence on the PAD prediction probability than the face mask and other facial regions (observations can be found in Section 4.6.2). 2) Fu *et al.* [90] explored the contributions of different face sub-regions to the face image quality and their experiments indicated that the eye region quality largely affects FR performance and additionally shows consistent quality degradation in face morphing attacks [91]. Such observations suggest that the eye regions comprise more subtle and discriminative information for various face-related tasks. As a result, we propose to weigh the predicted feature map regionally instead of just calculating the overall mean score. In our experiment, the weight of the eye region (including both the eyes and eyebrow region) is set to 0.6, while the weight for the face mask region is 0.1 and 0.3 for other regions, motivated by the observations in Section 4.6.2. Figure 4.4 illustrates the RW in the test phase, and the lighter color in the regional weight map refers to higher weight. The output feature map is multiplied with the regional weight map by the Hadamard product. Finally, a mean value of the weighted feature map is computed as the final decision score.

4.5. Experimental setup

This section presents the experimental setups and the implementation details applied in the work.

4.5.1. Face recognition algorithms

For FR systems, trained CNNs are typically used as feature extractors. The feature vector extracted from a specific layer of an off-the-shelf CNN was used as the template to represent the corresponding input face image. Then, the resulting templates were compared with each other using similarity measures. To provide a vulnerability analysis of the FR systems to our novel masked attacks, we adapted the following three FR algorithms:

- **ArcFace:** ArcFace [68] introduced an additive angular margin loss function to obtain highly discriminative features for FR. We chose this algorithm because ArcFace consistently outperformed state-of-the-art methods. ArcFace achieved 99.83% on Labeled Faces in the Wild (LFW) [118] and 98.02% on YouTube Faces (YTF) [235] dataset. The pre-trained ArcFace model [1] in our study was based on the ResNet-100 [107] architecture and trained on the MS-Celeb-1M [100] dataset (MS1M-v2). The output template is a 512-dimension feature vector extracted from the '*fc1*' layer of ArcFace.

- **SphereFace:** Liu *et al.* [167] proposed a deep hypersphere embedding approach (SphereFace) for FR task. SphereFace [167] utilized the angular softmax loss for CNNs to learn angularly discriminative features. This method also achieved competitive performance on LFW [118] (accuracy of 99.42%) and YTF [235] datasets (95.00%). We extract the face representation with 512-dimension from a pre-trained 20-layer SphereFace model. [2]

- **VGGFace2:** The first version of VGGFace is based on 16-layer VGG [207] network, while the second version of VGGFace (VGGFace2) [28] adopt ResNet-50 [107] as the backbone architecture. In this work, we use the second version that a ResNet-50 network trained on VGGFace2 dataset [28] [3] for extracting the 512-dimension templates.

The vulnerability of each FR system to attacks was analyzed based on three scenarios. Regardless of the scenario, the references are scenarios-specific bona fide videos captured on the first day, while bona fide videos from the second and third days or attack videos

[1] The official ArcFace model: `https://github.com/deepinsight/insightface`

[2] The official SphereFace model: `https://github.com/wy1iu/sphereface`

[3] The VGGFace2 model: `https://github.com/WeidiXie/Keras-VGGFace2-ResNet50`

were selected as probes. The three cases, including the division of scenario-specific references and probes, are described with the results in detail in Section 4.6.1. Once the references for the face images are obtained, we use the Cosine-similarity as recommended in [28, 68, 167] to compute the similarity scores between references and probes.

4.5.2. Face recognition experimental settings

The vulnerability of each FR system on each type of PA is analyzed based on three experimental settings. In the first setting BM0-BM0, we use the bona fide unmasked samples captured on the first day as references. Then, the references are compared against bona fide BM0 samples captured on the second and third days of the same subjects (to compute genuine scores), as well as of other subjects (zero-effort imposter (ZEI) scores). Once genuine and ZEI comparison scores are obtained, the operating threshold is computed using the $\tau_{FMR@0.01}$ threshold. Finally, the probe samples of each type of PA were compared against the reference of the same subjects separately. In the second setting BM0-BM1, the difference is that bona fide BM1 data captured on the second and third days are used for comparison against references BM0 and then obtain the corresponding genuine and ZEI scores. In the third setting, BM1-BM1, the bona fide masked faces captured on the first day are references for each subject. Such references are also compared against the masked bona fide samples captured on the second and third days to obtain their genuine and ZEI scores. These three experimental settings are provided to enable addressing the following four questions: 1) When having an unmasked reference and we use a decision threshold that does not consider masked comparisons (BM0-BM0), how vulnerable are FR systems to the three types of attacks in CRMA (AM0, AM1, and AM2)? 2) When having an unmasked reference and we use a decision threshold based on unmasked-to-masked comparisons (BM0-BM1), how vulnerable are FR systems to the three types of attacks in CRMA (AM0, AM1, and AM2)? 3) When having a masked reference and we use a decision threshold based on masked-to-masked comparisons (BM1-BM1), how vulnerable are FR systems to the three types of attacks in CRMA (AM0, AM1, and AM2)? Additionally, we address the fourth question: 4) will the vulnerability of FR systems be different when facing the AM1 and AM2 attacks?

4.5.3. Face recognition evaluation metrics

To measure the performance of FR techniques, the *genuine match rate* (GMR), which refers to the proportion of correctly matched genuine samples, is used at the fixed FMR. GMR is equal to 1 minus the FNMR. Moreover, to analyze the vulnerability of FR algorithms for our masked attacks, the IAPMR corresponding to the proportion of PAs accepted by the FR system as genuine presentations is adopted. IAPMR also follows the standard definition presented in ISO/IEC 30107-3 [127]. The threshold for GMR and IAPMR is defined by fixing the FMR at 1% (denoted as $\tau_{FMR@0.01}$). The probe images with similarity scores lower than the $\tau_{FMR@0.01}$ are not matched. Moreover, the recognition score-distribution histograms are shown in Figure 4.6, 4.7, and 4.8. In addition to these metrics, the EER value, where FMR equals FNMR, is computed to compare the FR algorithms.

4.5.4. Face PAD algorithms

A competition [13] was carried out in 2017 to evaluate and compare the generalization performance of face PAD techniques under real-world variations. In this competition [13], there were 14 participating teams together with organizers that contributed to several state-of-the-art approaches. We chose two methods (as previously discussed in Section 4.2), the LBP-based method (referred to as the baseline in [13]), and hybrid CPqD, and included additional solutions. We re-implemented a total of seven face PAD algorithms in this study, which can be categorized into three groups: hand-crafted features, deep-learning features, and hybrid features. For further cross-database evaluation scenarios, we used three publicly available databases, mainly involving 2D PAs (details in Section 4.2): CASIA-FAS [261], MSU-MFS [233], and OULU-NPU [14] in the competition. A brief description of the adopted methods is provided below:

- **LBP:** The LBP method is referred to as baseline method in [13] provided by the competition organizers that utilized the color texture technique. The face in a frame is first detected, cropped, and normalized to a size of 64×64 pixels. Second, an RGB face was converted into HSV and YCbCr color spaces. Third, the LBP features were extracted from each channel. The obtained six LBP features are then concatenated into one feature vector to feed into a softmax classifier. The final prediction score for each video was computed by averaging the output scores of all the frames.

- **CPqD:** The CPqD is based on the Inception-v3 network [213] and the above LBP method. The last layer of the pre-trained Inception-v3 model was replaced by a fully connected layer and a sigmoid activation function. The faces in the RGB frames are detected, cropped, and normalized to 299×299 pixels. These face images were utilized as inputs to fine-tune the Inception-v3 model. The model with the lowest EER on the development set among all 10 training epochs was selected. A single score for each video was obtained by averaging the output scores of all frames. To further improve the performance, the final score for each video was computed by fusing the score achieved by the Inception-v3 model and the score obtained by the LBP method.

- $\mathbf{Inception_{FT}}$ and $\mathbf{Inception_{TFS}}$: Since the CPqD uses the Inception-v3 [213] network as the basic architecture, we also report the results of fine-tuned Inception-v3 model, named $\mathrm{Inception}_{\mathrm{FT}}$. In addition to the fine-tuned model, we trained the Inception-v3 model from scratch for performance comparison, named $\mathrm{Inception}_{\mathrm{TFS}}$. In the training phase, the binary cross-entropy loss function and Adam optimizer with a learning rate of 10^{-5} were used. The output scores of the frames were averaged to obtain a final prediction decision for each video.

- $\mathbf{FASNet_{FT}}$ and $\mathbf{FASNet_{TFS}}$: FASNet [172] used transfer learning from pre-trained VGG16 model [207] for face PAD. They used a pre-trained VGG16 model as a feature extractor and modified the last fully connected layer. The newly added fully connected layers with a sigmoid function were then fine-tuned for the PAD task. This fine-tuned FASNet is referred to as $\mathrm{FASNet}_{\mathrm{FT}}$, similar to the Inception-v3 network methods, and we also train FASNet from scratch with the name $\mathrm{FASNet}_{\mathrm{TFS}}$. The input images are the detected, cropped, and normalized RGB face frames with a size of 224×224 pixels. The Adam optimizer with a learning rate of 10^{-4} was used for training, as defined in [172]. Data augmentation techniques and class weights are utilized to deal with imbalanced data problems. To further reduce overfitting, an early stop technique with a patience of 5 and maximum epochs of 30 was used. The resulting scores were averaged to obtain the final score for each video.

- **DeepPixBis:** George *et al.* [93] proposed a densely connected network framework for face PAD with binary and deep pixel-wise supervision. This framework is based on DenseNet architecture [117]. Two dense blocks and two transition blocks with a fully connected layer with sigmoid activation produce a binary output. We used the same data augmentation technique (horizontal flip, random jitter in brightness, contrast, and saturation) and the same hyper-parameters (Adam optimizer with a learning rate of 10^{-4} and weight decay of 10^{-5}) as defined in [93] for the training. In addition to data augmentation, we applied the class weight and an early stopping technique to avoid overfitting. The final score for each video was computed by averaging the frame scores.

4.5.5. PAD experimental protocols

To study the possible effect of face masks and spoofing faces with real masks on the performance of PAD and FR systems, we provide three protocols for further experimental analysis.

PAD protocols for the CRMA database

In this study, three protocols are provided to study the impact of masks on the performance of PAD solutions under different training settings. Other factors, such as various devices, illumination, and capture scales, are outside the scope of this study. These three protocols try to answer three questions separately: 1) Does the PAD algorithm trained on unmasked data generalize well on the masked bona fides and attacks, that is, can the previously trained model be adapted to the present-day situation? 2) Does the PAD algorithm designed before the COVID-19 pandemic still work efficiently if it is trained on additional masked data? 3) Will a network that has learned masked face attacks be confused by real masks that obscure the spoof face? Hence, we split 47 subjects in the CRMA database into three subject-disjoint sets: the training set (19 subjects), the development set (10 subjects), and the testing set (18 subjects). Gender was balanced as much as possible between the three sets. Table 4.2 provides more information about three protocols. A detailed description of three protocols is as follows:

Protocol-1 (P1): This protocol demonstrates the generalization performance of the PAD solutions trained on unmasked data. The training and development sets contain only videos of subjects without masks (such as data in most current PAD databases). The trained model was then tested on the data using face masks. More specifically, only BM0 and AM0 data were used for training, while BM1, AM1, and AM2 were considered unknown mask data.

Protocol-2 (P2): In contrast to protocol-1, which focuses on generalizability on unseen mask data, the second protocol is designed to evaluate the performance of PAD algorithms when masked data has been learned in the training phase. In this protocol, the training, development, and testing sets include masked and unmasked bona fides (BM0, BM1), masked and unmasked attacks (AM0, AM1), and spoof faces with real masks (AM2).

Protocol-3 (P3): Until now, the effect of AM2 on PAD performance is still unclear. AM2 is a special attack type that a real face mask is placed on spoof faces, which means it contains only partial artifacts (i.e., unmasked face spoofing region) compared to AM1, which carries entire artifacts (i.e., spoofed face and mask). Therefore, this protocol attempts to answer the following question: If the network has learned the masked attacks AM1, can this trained model not be confused by a real mask and perform similarly on the attack covered by a real mask AM2? Consequently, the training and development sets include bona fides BM0 and BM1, and attacks AM0 and AM1, while AM2 is an unknown attack in the testing set. Because data in the CRMA are video sequences and the number of videos between bona fide and attack classes are imbalanced, we sampled 60 frames from a bona fide video and five frames from an attack video to reduce data bias. In addition to different frame sampling, we also adapt the class weights inversely proportional to the class frequencies to reduce overfitting in the training phase. In the test phase, a final classification decision was determined by averaging the prediction scores of all sampled frames.

Protocol	Set	Subjects	Types of masks	# BF videos	# Attack videos
P1	Train	1-19	BM0, AM0	57	1569
	Dev	20-29	BM0, AM0	30	810
	Test	30-47	BM0, BM1, AM0, AM1, AM2	162	4860
P2	Train	1-19	BM0, BM1, AM0, AM1, AM2	171	5130
	Dev	20-29	BM0, BM1, AM0, AM1, AM2	90	270
	Test	30-47	BM0, BM1, AM0, AM1, AM2	162	4860
P3	Train	1-19	BM0, BM1, AM0, AM1	171	4617
	Dev	20-29	BM0, BM1, AM0, AM1	90	2430
	Test	30-47	BM0, BM1, AM0, AM1, AM2	162	4860

Table 4.2.: The detailed information of three protocols for exploration of the possible effect of face masks. The bona fide is denoted as BF. The test data is the same in the three protocols, while the types of training and development data are different.

PAD protocols for cross-database scenarios

In addition to the intra-database scenario on our CRMA database, we also perform cross-database experiments to explore the generalizability of these PAD algorithms on masked data. Because the PAIs in the CRMA database are print and replay attacks, we selected three popular publicly available databases containing the same PAIs: CASIA-MFS [261], MSU-MFS [233], and OULU-NPU [14] to demonstrate the evaluation. We conducted two cross-database experiments. In the first cross-database scenario, the PAD solutions trained on three publicly available databases were evaluated on the test set of the CRMA database. In addition, the results tested on their own test sets are also reported (as shown in the left block in Table 4.5). The first setting is similar to protocol-1 of the CRMA intra-database scenario, as no masked data are seen in the training phase. Therefore, the first cross-database setting is also used to answer the first question: does the PAD algorithm trained on unmasked data generalize well on masked bona fides and attacks? Conversely, in the second cross-database experiment, models trained on different protocols of the CRMA database were evaluated separately on publicly available databases. This experimental setting can help us understand the CRMA database values beyond face masks, such as the diversity of masks/sensors/scales. However, the second scenario does not support the main study of the work and is provided only for completeness; thus, the results are reported in the supplementary material. In both cross-database scenarios, we use the

$\tau_{BPCER10}$ decision threshold computed on the development set of the training database as a priori to determine the APCER, BPCER, and HTER values of the test database.

4.5.6. PAD evaluation metrics

The metrics following the ISO/IEC 30107-3 [127] standard were used to measure the performance of the PAD algorithms: APCER and BPCER as introduced in Chapter 2. The APCER and BPCER reported in the test set were based on a pre-computed threshold in the development set. In our study, we use a BPCER at 10% (on the development set) to obtain the threshold (denoted as $\tau_{BPCER10}$). Additionally, HTER is used for the cross-database evaluation. Noticeably, we computed a threshold in the development set of the training database. Then, this threshold was used to determine the HTER value in the test database. The Detection Equal Error Rate (D-EER) (D-EER) value, where APCER and BPCER are equal is also reported in the cross-database scenarios. For further analysis of PAD performance, ROC curves were also demonstrated.

4.5.7. Model training setup

We following the implementation details described in [18, 93]. First, a face was detected and cropped by MTCNN [256]. Then, the 68 facial landmarks were detected by Dlib library [149]. Based on the jaw landmarks, a pixel-wise label was generated for the partial attack (AM2), in which the mask area is labeled as bona fide, and the rest of the face is labeled as an attack. For bona fide samples (BM0 and BM1), the values in this binary mask are all set to one (bona fide), while they are all set to zero for the other attack samples (AM0 and AM1). Finally, the input face image and the generated binary mask were used jointly while applying augmentation and resized to $224 \times 224 \times 3$ pixels and 14×14 pixels, respectively. In the experimental result discussion, we use PAL-RW$_{DeepPixBis}$ and PAL-RW$_{MixFaceNet}$ to indicate proposed PAL-RW solution by utilizing DeepPixBis and MixFaceNet network backbones, respectively. For PAL-RW$_{DeepPixBis}$, the same augmentation techniques (horizontal flip and random jitter with the probability of 0.5) of DeepPixBis [93] were used in the training phase. The combined training loss was minimized by Adam Optimizer with the learning rate of 10^{-4} and the weight decay of 10^{-5}. For PAL-RW$_{MixFaceNet}$, we employed the above augmentation techniques and the SGD optimizer with the learning rate of 10^{-2} and the weight decay of 5^{-3}. The Exponential learning rate scheduler was adopted with the gamma of 0.995. To further avoid overfitting, we applied class weight due to the unbalanced data and an early stopping technique with the maximum epoch of 100 and the stop patience of 15 for both models training processes. In the test phase, a regional weight map was generated based on the detected facial landmarks. The prediction score of each frame was a mean value of the regional weighted feature map. Finally, a final score for each video was computed by averaging the scores of the processed frames.

4.6. Results

4.6.1. Face recognition vulnerability

The performance and vulnerability of each FR system are summarized in Table 4.3. SphereFace [167] obtains relatively low IAPMR values; however, its GMR values are also much lower than those of ArcFace [68] and VGGFace [28]. In general, the IAPMR values of all three FR systems were close to their GMR values. Specifically, FR systems are vulnerable to unmasked attacks when unmasked bona fide samples are used as references (the settings BM0-BM0 and BM0-BM1), and vulnerable to the masked attacks when the reference is masked bona fide data. Comparing the vulnerability analysis results for AM1 and AM2 in all three cases and all FR systems, we note that the IAPMR values of AM2 are always significantly higher than those of AM1. This indicates that applying real masks on attack presentations can further reduce the performance of FR systems. This might be due to the fact that the AM2 attacks possess more realistic features than AM1.

Settings	Attack Probes	ArcFace[68]			SphereFace [167]			VGGFace [28]		
		EER	GMR	IAPMR	EER	GMR	IAPMR	EER	GMR	IAPMR
BM0 - BM0	AM0	0.00	100	98.40 [98.22, 98.56]	8.57	75.85	66.31 [65.69, 66.93]	0.12	100	99.47 [99.37, 99.56]
	AM1			81.61 [81.24, 81.97]			2.80 [2.65, 2.96]			71.54 [71.12, 71.96]
	AM2			97.10 [96.77, 97.41]			10.45 [9.89, 11.03]			97.23 [96.91, 97.53]
BM0 - BM1	AM0	2.25	96.56	98.73 [98.58, 98.88]	22.83	19.99	84.17 [83.68, 84.64]	2.29	94.2	99.86 [99.80, 99.90]
	AM1			88.57 [88.27, 88.86]			15.26 [14.92, 15.60]			90.24 [89.96, 90.51]
	AM2			98.56 [98.33, 98.78]			40.00 [39.09, 40.91]			99.55 [99.41, 99.67]
BM1 - BM1	AM0	1.00	99.00	70.62 [70.19, 71.04]	13.13	59.33	2.43 [2.29, 2.58]	0.85	99.46	45.84 [45.38, 46.31]
	AM1			94.20 [94.04, 94.35]			47.69 [47.36, 48.02]			97.41 [97.30, 97.51]
	AM2			97.70 [97.49, 97.89]			50.82 [50.16, 51.48]			98.26 [98.08, 98.43]

Table 4.3.: The performance and vulnerability of FR systems. The GMR and IAPMR values were computed based on the $\tau_{FMR@0.01}$ threshold. The 95% confidence intervals for the IAPMR values are shown in parentheses.

To further verify this assumption, we provide histograms of the similarity score distribution in the three scenarios and three FR systems (see Figure 4.6, 4.7, and 4.8). In the histograms, green refers to genuine scores, blue represents ZEI scores, and gray represents attack verification scores. The ideal situation is that there is no overlap between the green and the other two histograms. Figure 4.6 shows the score distributions of ArcFace [68], where the rows from top to bottom represent BM0-BM0, BM0-BM1, BM1-BM1 cases and columns from left to right refer to AM0, AM1, and AM2 attacks. It can be seen that 1) the verification scores of attacks are higher than the scores of ZEI in all cases. 2) The scores of AM0 attacks and genuine scores almost overlap in the BM0-BM0 and BM0-BM1 settings, while the scores of AM1/AM2 attacks have many overlapping areas with genuine scores in the BM1-BM1 setting. 3) for all cases, the scores of AM2 have more overlaps with genuine scores than AM1. Similar observations can be found in Figure 4.7 for the SphereFace, and Figure 4.8 for VGGFace. These observations are consistent with the findings presented in Table 4.3.

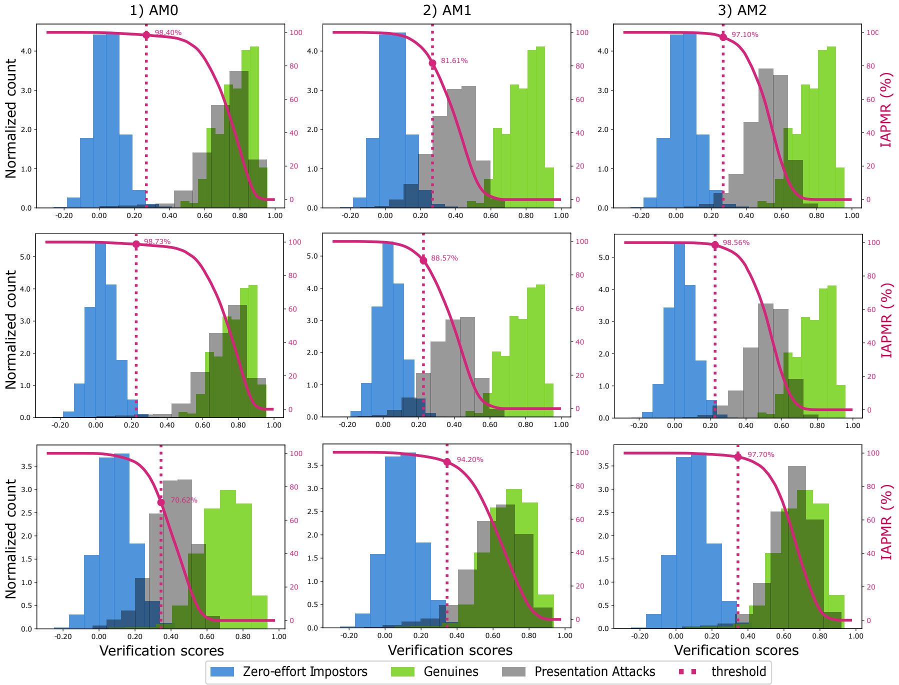


Figure 4.6.: The similarity score distributions by off-the-shelf ArcFace [68]. The rows from top to bottom represent three experimental settings: BM0-BM0, BM0-BM1, BM1-BM1, as shown in Table 4.3.

Overall, these results indicate that: 1) FR systems are more vulnerable to unmasked attacks compare to masked attacks when the references are unmasked faces; 2) When the threshold is computed based on the unmasked-to-masked comparison, the vulnerability of FR systems becomes higher for both masked or unmasked attacks; 3) When the reference is masked, FR systems are more vulnerable to masked attacks in comparison to the FR systems having unmasked references. Another important finding is that 4) FR systems

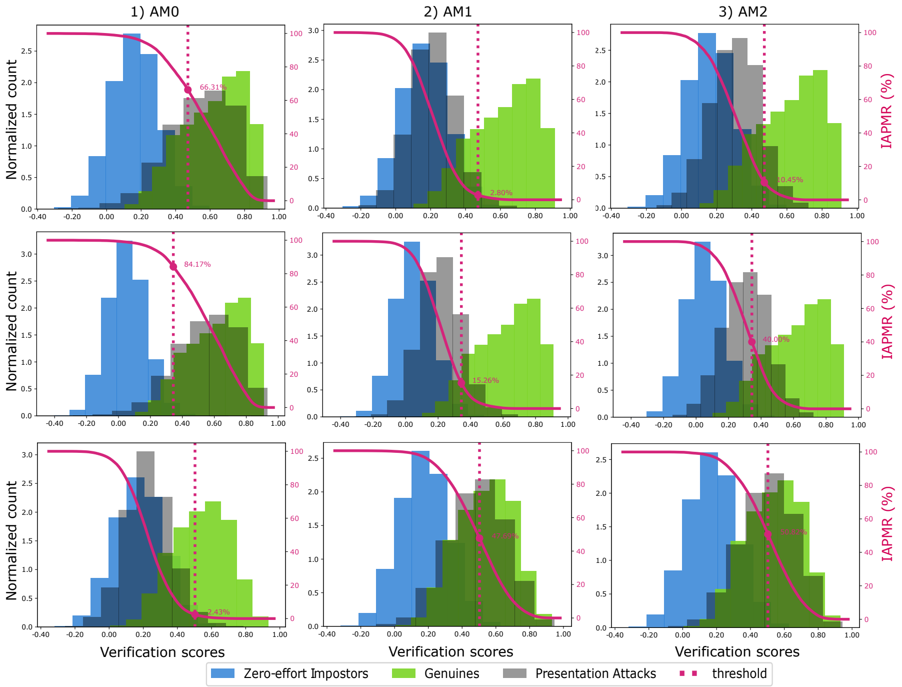


Figure 4.7.: The similarity score distributions by off-the-shelf SphereFace [167].

pose a higher vulnerability for spoof faces with real masks placed on them (AM2) than a masked face attack (AM1). Such observations raise concerns about the security of FR systems when facing masked attacks.

4.6.2. Quantitative and qualitative analysis of three PAD protocols

Analysis of Protocol-1

Protocol-1 represents the pre-COVID-19 PAD scenarios, in which subjects normally do not wear a mask, and demonstrates the generalization performance on masked data. Therefore, protocol-1 is considered the most challenging task because of the unseen BM1, AM1, and AM2 data. Table 4.4 describes the results of the different protocols on the CRMA database. The bold numbers indicate the highest BPCER values between BM0 and BM1 and the highest APCER values between AM0, AM1, and AM2 in each PAI. By observing the first block, P1, in Table 4.4, the BPCER values of masked bona fide samples are much higher than those of unmasked ones; however, most PAD systems achieve lower APCER values on the masked attack samples (either AM1 or AM2). The higher classification error rates on

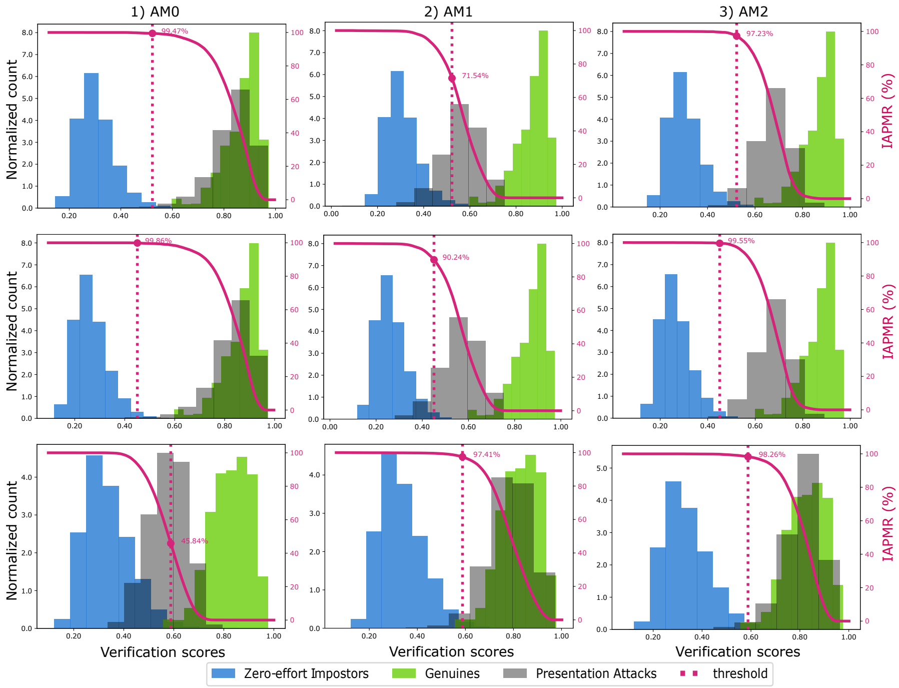
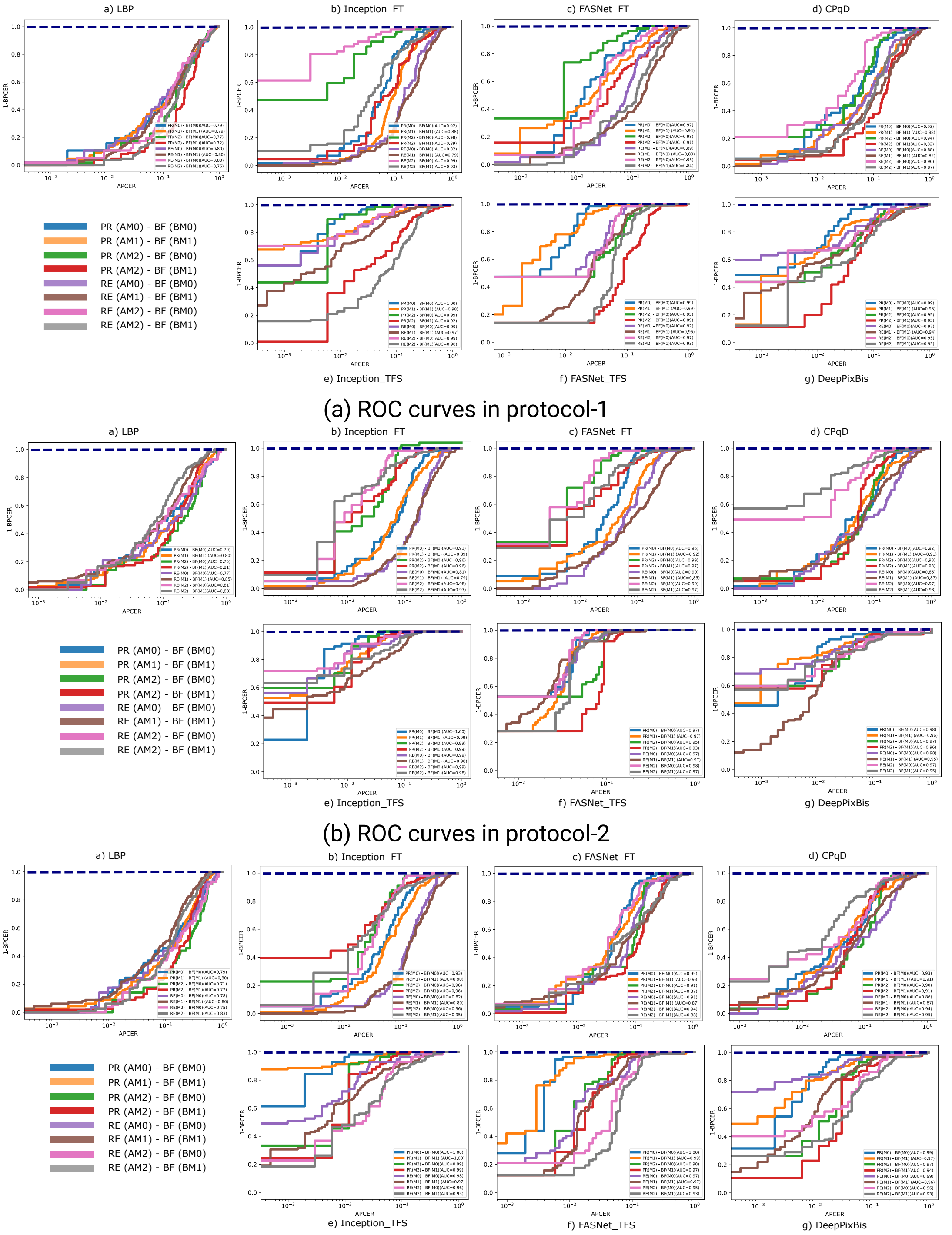


Figure 4.8.: The similarity score distributions by off-the-shelf VGGFace [167].

masked bona fide and the lower error rates on masked attacks are intuitively conceivable. When the model has not seen faces wearing a mask before, it is more inclined to falsely classify such a masked bona fide sample (BM1) as an attack. Moreover, it is interesting to note that networks trained from scratch and the DeepPixBis approach work worse on attack AM2 than AM1. These observations are consistent with the ROC (Figure 4.9). The red curves generated by printed AM2, bona fide BM1, and gray curves obtained by replay AM2 and bona fide BM1 possess significantly smaller areas under the curves in five of the seven methods. Furthermore, training a network from scratch improves the overall performance. The possible reason for those observations is that learning from scratch is more efficient for obtaining discriminative features between bona fide and artifacts. On the contrary, such approaches might be confusing when applying realistic masks to attack samples.

In addition to the intra-database scenario, the first cross-database experiment (introduced in Section 4.5.5) can be seen as similar to protocol-1, as both scenarios study PAD methods that PAD solutions trained on unmasked data and tested on the CRMA database. In Table 4.5, the bold BPCER number is the highest BPCER (between BM0 and BM1) for



(a) ROC curves in protocol-1

(b) ROC curves in protocol-2

(c) ROC curves in protocol-3

Figure 4.9.: ROC curves for all PAD methods in three protocols. Eight combinations between bona fide and attack (testing data) are represented for each method in each protocol: PR(AM0)-BF(BM0), PR(AM1)-BF(BM1), PR(AM2)-BF(BM0), PR(AM2)-BF(BM1) in print PAI and RE(AM0)-BF(BM0), RE(AM1)-BF(BM1), RE(AM2)-BF(BM0), RE(AM2)-BF(BM1) in replay PAI. The x-axis and y-axis are APCER and 1 - BPCER, respectively. The red curves (PR(AM2)-BF(BM1)) and gray curves (RE(AM2)-BF(BM1)) show significantly smaller AUC values by most PAD methods on protocol-1. Moreover, $\mathrm{Inception}_{\mathrm{TFS}}$, $\mathrm{FASNet}_{\mathrm{TFS}}$, and DeepPixBis achieve higher AUC values on protocol-2 and -3 than on protocol-1 might be due to the masked data in the training phase.

Protocol	Method	Threshold @ BPCER 10% in dev set							
		BPCER (%)		APCER (print) (%)			APCER (replay) (%)		
		BM0	BM1	AM0	AM1	AM2	AM0	AM1	AM2
P1	LBP	1.75	**4.39**	**80.12**	72.61	71.93	**74.95**	67.76	73.98
	Inception$_{\text{FT}}$	19.30	**84.21**	**10.33**	3.80	2.92	**27.19**	5.81	0.88
	CPqD	7.02	**47.37**	**18.52**	7.80	15.79	**31.77**	11.19	10.23
	FASNet$_{\text{FT}}$	12.28	**56.14**	**7.02**	1.36	2.92	**20.37**	12.21	9.65
	Inception$_{\text{TFS}}$	7.04	**48.25**	1.36	0.00	**1.75**	**7.50**	0.34	7.02
	FASNet$_{\text{TFS}}$	7.02	**29.82**	1.95	0.49	**15.20**	**8.09**	4.64	7.89
	DeepPixBis	19.30	**28.95**	1.56	1.56	**5.85**	3.61	4.05	**6.43**
P2	LBP	**26.32**	11.40	31.38	**44.44**	36.84	**36.74**	34.39	28.95
	Inception$_{\text{FT}}$	1.75	**7.02**	**35.28**	30.80	11.70	**54.09**	52.17	10.23
	CPqD	3.51	**7.89**	27.49	**30.41**	16.37	**46.20**	44.50	10.23
	FASNet$_{\text{FT}}$	1.75	**17.54**	10.72	**12.77**	5.85	**30.60**	28.09	3.80
	Inception$_{\text{TFS}}$	8.77	**18.42**	0.78	1.56	**2.34**	3.90	**5.23**	2.63
	FASNet$_{\text{TFS}}$	14.04	**29.82**	4.09	3.41	**9.36**	**4.69**	2.88	3.80
	DeepPixBis	**29.82**	24.56	0.78	0.19	**1.75**	0.10	**1.86**	0.88
P3	LBP	**22.81**	9.65	35.28	**48.15**	47.95	38.50	36.79	**42.40**
	Inception$_{\text{FT}}$	1.75	**8.77**	24.17	**24.37**	11.70	46.69	**47.14**	14.04
	CPqD	**7.02**	**7.02**	20.66	**28.95**	21.64	41.23	**41.52**	17.84
	FASNet$_{\text{FT}}$	5.26	**21.93**	14.04	9.94	**26.71**	**22.62**	19.88	20.47
	Inception$_{\text{TFS}}$	21.05	**21.93**	0.19	0.00	**1.17**	1.56	2.34	**4.97**
	FASNet$_{\text{TFS}}$	22.81	**34.21**	0.39	0.29	**2.34**	3.41	2.20	**6.43**
	DeepPixBis	17.54	**24.56**	0.78	0.68	**2.92**	0.88	1.91	**6.43**

Table 4.4.: The PAD performance of different PAD solutions in three protocols (as described in Section4.5.5). The bold number in each protocol and each method refers to the highest BPCER on BM0 and BM1 data and the highest APCER value between AM0, AM1, and AM2 in the two PAIs, respectively. The higher BPCER values for BM1 (in comparison to BM0) indicate that subjects wearing masks tend to be classified falsely as attacks.

each PAD method. The bold APCER number is the highest APCER (between AM0, AM1, and BM2) for each PAD method in print and replay attacks, respectively. This bolding is performed to show which samples are more difficult to classify correctly. We observed that the performance in the cross-database setting was relatively poor for all models. Even though deep-learning-based methods achieved great results on their own test sets, they generalize significantly worse on masked bona fide samples; for example, most BPCER values for BM1 are close to 100%. In contrast, most algorithms achieve lower APCER values on masked AM1 and AM2 than unmasked AM0 attacks, which is consistent with the observation of protocol-1 from the intra-database scenarios.

Trained on	Method	Threshold @ BPCER 10% in dev set of trained database											
		Tested on the same dataset (%)			Tested on our CRMA dataset (%)								
		D-EER	BPCER	APCER	BPCER		APCER (Print)			APCER (Replay)			
					BM0	BM1	AM0	AM1	AM2	AM0	AM1	AM2	
CAISA-FASD	LBP	7.50	6.25	8.75	38.60	**56.14**	**42.11**	24.76	18.13	**60.72**	34.59	22.51	
	Inception$_{\mathrm{FT}}$	10.00	8.75	15.00	21.05	**38.60**	**35.48**	5.95	16.96	**69.49**	47.44	15.50	
	CPqD	6.25	11.25	3.12	38.60	**65.79**	**31.97**	12.38	8.77	**53.22**	23.06	14.62	
	FASNet$_{\mathrm{FT}}$	8.75	12.50	4.38	15.79	**90.35**	**44.83**	2.14	23.98	**64.13**	5.76	22.81	
	Inception$_{\mathrm{TFS}}$	*0.00*	*1.25*	*0.00*	12.28	**20.08**	**61.60**	40.35	49.71	**90.35**	83.19	59.65	
	FASNet$_{\mathrm{TFS}}$	1.25	3.75	0.62	21.05	**75.44**	**60.23**	19.49	38.60	**70.86**	16.32	45.61	
	DeepPixBis	1.25	6.25	*0.00*	35.09	**66.67**	**70.57**	36.65	56.73	**57.99**	29.26	42.98	
MSU-MFSD	LBP	4.17	*4.17*	4.17	98.25	**100.00**	0.58	**0.68**	0.00	**3.22**	2.25	0.00	
	Inception$_{\mathrm{FT}}$	20.14	20.81	16.67	**50.88**	25.44	47.95	**56.04**	52.05	31.19	**48.85**	44.15	
	CPqD	4.17	*4.17*	4.17	98.25	**100.00**	0.19	**0.39**	0.00	**1.46**	1.56	0.00	
	FASNet$_{\mathrm{FT}}$	13.19	26.39	4.17	43.86	**85.96**	**32.55**	2.63	0.58	**42.50**	13.39	2.34	
	Inception$_{\mathrm{TFS}}$	4.17	8.33	1.39	80.70	**94.74**	**0.19**	0.00	0.00	**8.58**	0.78	2.05	
	FASNet$_{\mathrm{TFS}}$	*0.00*	8.44	*0.00*	91.23	**100.00**	0.00	0.00	0.00	**7.70**	0.00	0.29	
	DeepPixBis	*0.00*	*4.17*	*0.00*	**82.46**	80.70	0.00	**0.10**	0.00	10.33	**10.36**	5.26	
Oulu-NPU	LBP	8.33	7.50	10.21	40.35	**67.54**	**35.28**	25.54	13.45	**26.12**	10.89	13.74	
	Inception$_{\mathrm{FT}}$	15.00	16.67	11.04	61.40	**87.72**	**11.50**	5.85	8.77	**12.38**	2.39	1.46	
	CPqD	8.33	9.17	3.54	57.89	**89.47**	**9.55**	3.70	1.17	**10.14**	1.03	0.58	
	FASNet$_{\mathrm{FT}}$	3.23	*1.67*	4.38	49.12	**73.68**	**33.92**	27.10	8.77	**22.81**	8.99	3.80	
	Inception$_{\mathrm{TFS}}$	4.17	3.33	6.46	80.07	**100.00**	**22.81**	0.78	2.34	**3.22**	0.00	0.00	
	FASNet$_{\mathrm{TFS}}$	5.10	11.67	3.33	70.18	**99.12**	**46.98**	18.03	19.88	**8.09**	0.39	0.29	
	DeepPixBis	*2.29*	2.92	*0.00*	66.67	**98.25**	**44.64**	11.21	4.68	**10.23**	0.10	0.58	

Table 4.5.: Cross-database evaluation 1: the model trained on three publicly available databases is used to test on the CRMA database. This cross-database scenario is similar to protocol-1, as no masked data is seen during the training phase. Italic numbers indicate the lowest error rate on their own test set, and bold numbers indicate the highest error rate in the bona fide and each PAI category. The results show that despite good performance on their own test set, these trained models do not generalize well to masked bona fides and attacks.

In general, the experimental results of the intra-database protocol-1 and the first cross-database scenario results answer the first posed question (in Section 4.5.5) by showing that models trained only on unmasked data cannot properly classify images of masked faces. A subject with a mask on has a high probability of being falsely detected as an attack by PAD systems, even if this subject is bona fide.

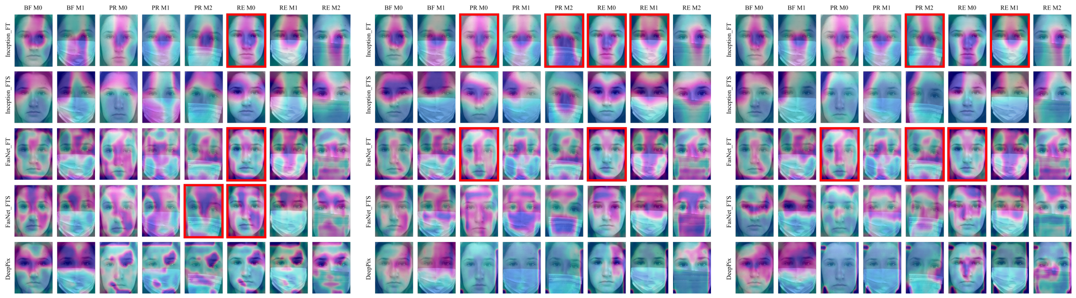
Analysis of Protocol-2

Protocol-2 targets the performance of PAD algorithms on masked data when both unmasked and masked samples are used in the training phase. As shown in Table 4.4, we can observe the following points: First, despite the fact that the masked bona fide samples are still more difficult to classify correctly than unmasked ones in most cases, the BPCER value of BM1 behaves more similar to its behavior on the BM0 in protocol-2 than in protocol-1. Moreover, the BPCER values of BM0 and BM1 in protocol-2 decreased in most cases compared with the results of protocol-1. For example, the BPCER value of BM1 achieved by $\mathrm{Inception}_{\mathrm{FT}}$ was 84.21% in protocol-1 and 7.02% in protocol-2. This finding indicates that learning the masked data is helpful in improving the performance of PAD methods. This is also consistent with the observation in the ROC curves (by comparing the ROCs in protocol-1 and protocol-2 in general). In particular, $\mathrm{Inception}_{\mathrm{TFS}}$, $\mathrm{FASNet}_{\mathrm{TFS}}$, and DeepPixBis achieved significant progress (larger areas under the curves). Second, six of the seven methods performed worse on the masked printed face (AM1 or AM2), while five of the seven algorithms showed inferior results for unmasked replay attacks. Moreover, AM2 in print PAI achieves higher APCER values than AM1 by training from scratch approaches. One possible reason for the different results between print and replay attacks is specular reflection. Because attack data were collected in windowless labor with all electric lights on, tablets easily reflect the light compared to the printed paper, and this reflection is difficult to avoid. The real face masks might also leak light when placed on an electric tablet, but this does not appear when applied on printed paper. In general, the experimental results of the intra-database protocol-2 answer the second question (in Section 4.5.5), which addresses the performance changes of the current PAD algorithms after complementary learning on the masked data. Based on the above findings, we can conclude that the PAD algorithms still perform worse on masked bona fides (BM1) than on unmasked faces (BM0), even when the PAD solutions are trained on masked data.

Analysis of Protocol-3

Protocol-3 investigates the generalizability of the model trained on data that includes masked face attacks (AM1) when tested on the masked face attacks where a real mask is placed on top of the attack (AM2). For bona fide samples, we draw a similar conclusion to protocol-1 and protocol-2, stating that masked bona fide samples have a higher probability of incorrectly being classified as attacks. However, the experimental results show differences in attack detection behavior (APCER) between protocol-3 on one side and protocols-1 and -2 on the other side. In this protocol, the highest APCER values of most PAD algorithms appear on either the AM1 or AM2 attacks in both print and replay PAIs. Second, the traditional LBP method, $\mathrm{Inception}_{\mathrm{FT}}$, $\mathrm{FASNet}_{\mathrm{FT}}$, and the hybrid CPqD method that achieve relatively worse results on AM0 or AM1 attacks than other methods may have proved to be unable to learn or extract sufficient discriminative features. Third, although the other methods, such as learning from scratch $\mathrm{Inception}_{\mathrm{TFS}}$ and $\mathrm{FASNet}_{\mathrm{TFS}}$ or custom designed DeepPixBis achieve impressive results on seen AM0 and AM1 attacks, they generalize not well on unseen AM2 attacks. These observations answer the third question stated in Section 4.5.5 by stating that a network trained on masked face attacks (AM1) tends to produce confusing decisions on AM2, where a real mask is placed on an attack face.

Qualitative Analysis and Visualization



(a) Protocol-1 (b) Protocol-2 (c) Protocol-3

Figure 4.10.: Examples for attention maps generated by ScoreCAM of different PAD algorithms and different protocols. The rows from top to bottom in each protocol correspond to $\mathrm{Inception_{FT}}$, $\mathrm{Inception_{TFS}}$, $\mathrm{FASNet_{FT}}$, $\mathrm{FASNet_{TFS}}$, and DeepPixBis. The columns from left to right in each protocol refer to BM0, BM1, PR-AM0, PR-AM1, PR-AM2, RE-AM0, RE-AM1, RE-AM2. Faces with red boxes are misclassified.

To qualitatively analyze and interpret the deep-learning-based methods, the score-Weighted CAM [226] technique was adopted to localize the discriminative areas in face images. The rows from top to bottom correspond to $\mathrm{Inception_{FT}}$, $\mathrm{Inception_{TFS}}$, $\mathrm{FASNet_{FT}}$, $\mathrm{FASNet_{TFS}}$ and DeepPixBis. Figure 4.10a shows the results of protocol-1 (the example subject is in the test set). $\mathrm{Inception_{FT}}$ mainly focuses on the nose, including nearby parts of the masks, whereas $\mathrm{Inception_{TFS}}$ pays more attention to the upper region of the face. Similarly, $\mathrm{FASNet_{TFS}}$ reduces the attention paid to the masks and increases the concentration around the forehead. DeepPixBis concentrates around the eyes for both unmasked (BM0) and masked (BM1) bona fides. However, for attack samples, attention seems to be focused on the left eye and partial masks. In general, masks are noticed by all networks. The results of protocol-2 and protocol-3 for the same subjects are shown in Figure 4.10b, and Figure 4.10c. We noticed that 1) the attention areas of fine-tuned networks hardly change in the three protocols because of the fixed weights of layers before the last classification layer. 2) $\mathrm{Inception_{TFS}}$ in protocol-2 appears to focus on the upper face, including many more eye regions than in protocol-1. 3) $\mathrm{FASNet}_{\mathrm{TFS}}$ in protocol-2 concentrates much more on applied real masks than in protocol-3 where training without AM2. 4) DeepPixBis still works well on bona fide, but for attack samples, its attention seems to be distracted to the edge of images. Although DeepPixBis produces correct decisions, this observation raises a serious concern about its reliability and generalizability. This concern was confirmed by the cross-database evaluation. DeepPixBis generally obtains worse cross-database results than the other two training from scratch networks (details see Table 4.5). Finally, looking at attention maps in all protocols for this identity, we notice that except for the misclassified samples (with red boxes) that appear on print/replay AM0, print AM2 attacks are more easily to be incorrectly detected as bona fide than AM1 attacks. This finding is in line with the previous quantitative evaluation that AM2 attacks may confuse the PAD, even if the network has been trained by masked face attacks.

4.6.3. Ablation study and results of PAL-RW

The above subsections provide a comprehensive analysis on the performance of face PAD algorithms under various masked face scenarios. The experiment results indicate that PAD algorithms have a high possibility of detecting masked bona fide samples as attackers. This section will present the results of the proposed PAL-RW method based on the Protocol-2, which is designed to evaluate the performance of PAD algorithms when masked data has been learned in the training phase. First, the contribution of each module will be introduced in Table 5.4 and then the overall PAD performance in Table 4.7 will be discussed.

Ablation study in module PAL and RW

To further validate the usefulness of each component of the proposed PAL-RW solution, we conduct the experiments by gradually adding PAL and RW modules. The results are shown in Table 5.4. We report the APCER and BPCER values by using the threshold $\tau^{unmasked}_{BPCER10}$ that is pre-computed on only unmasked data in the development to build a realistic ablation study where the behavior of the PAD on masked data is still unknown. As shown in Table 5.4, adding any one of the two components (PAL or RW) to the backbones does improve the PAD performance (considering at the overall performance index metric ACER). Moreover, the contribution made by the RW component is more significant than PAL while using DeepPixBis [93] as a backbone. For example, the ACER value is reduced from 29.47% obtained by the DeepPixBis backbone to 18.58% achieved by adding the RW component. Note that both setups use the same trained model, only the output prediction maps are weighted differently. This finding confirms our assumption that stresses the importance

Backbone	RW	PAL	BPCER (%)		APCER (Print) (%)			APCER (Replay) (%)			ACER (%)
			BM0	BM1	AM0	AM1	AM2	AM0	AM1	AM2	
DeepPixBis			63.16	64.04	0.00	0.00	0.00	0.00	0.64	0.00	29.47
DeepPixBis	√		35.09	41.23	0.00	0.10	1.17	0.19	1.95	0.58	18.58
DeepPixBis		√	42.11	51.75	0.00	0.00	0.00	0.00	0.44	0.00	23.35
DeepPixBis	√	√	26.32	29.82	0.00	0.19	1.17	0.00	1.32	0.29	14.81
MixFaceNet			5.26	7.89	12.09	9.75	11.70	21.83	17.68	14.33	13.80
MixFaceNet	√		5.26	7.89	11.50	9.36	11.70	21.64	17.15	14.33	13.75
MixFaceNet		√	12.28	7.89	4.29	4.39	8.19	17.54	12.36	9.06	13.10
MixFaceNet	√	√	8.77	8.77	4.09	6.24	7.02	22.03	17.34	17.54	12.00

Table 4.6.: The PAD performance of the different step-wise ablation experiments using DeepPixBis and MixFaceNet as backbone network architectures on the CRMA database. Here, the used PAD decision threshold is the one scoring a BPCER 10% on only unmasked data in the development set. The results show the individual and joint benefits of our PAL and RW components on PAD performance.

of eye regions for PAD decisions. When using MixFaceNet [18] as the backbone, the reduction in the classification error rates is slightly smaller than DeepPixBis. Nevertheless, the basic MixFaceNet achieves lower overall PAD performance (13.80% ACER value) in comparison to the DeepPixBis with additional PAL and RW modules (14.81% ACER value). Such results indicate that MixFaceNet architecture possesses not only lower computational complexity but also higher generalization ability. In addition, the ROC curves for the ablation experiments are shown in Figure 4.11. The red curve (DeepPixBis-RW-PAL) and the grey curve (DeepPixBis-RW-PAL) are on top of other curves. Overall, the PAL and RW components can both improve the PAD performance in the CRMA database.

Comparison with established PAD solutions

Table 4.7 presents the results of the different investigated methods on the CRMA database of Protocol-2 in Table 4.4 and aims to put the performance achieved by the proposed PAL-RW solution in the perspective of the performance of established PAD solutions. The last two rows, PAL-RW$_{DeepPixBis}$ and PAL-RW$_{MixFaceNet}$, are results achieved by our PAL-RW methods. The other rows are the results of established PAD solutions earlier reported in Table 4.4. The BPCER and APCER values in Table 4.7 are determined by the threshold $\tau^{all}_{BPCER10}$. The bold number is the lowest ACER value indicating the best overall performance. As shown in Table 4.7, our PAL-WR method improves the overall PAD performance. For example, the ACER value decreases from the 13.13% achieved

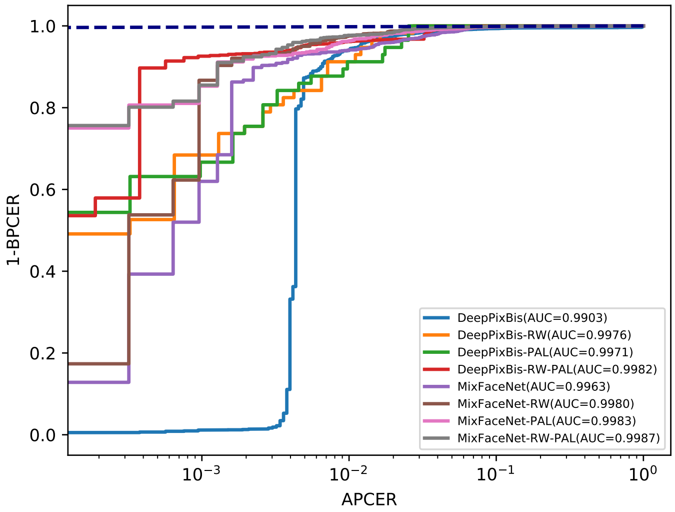


Figure 4.11.: ROCs of different solution steps of the ablation experiments in the CRMA database. Adding any one of the proposed modules (PAL or RW) improves the overall PAD performance. Moreover, experiments using the MixFaceNet backbone exhibit higher AUC values than using DeepPixBis architecture.

by DeepPixBis to 12.99% obtained by PAL-RW$_{DeepPixBis}$. In addition, the APCER value of the partial print attack (AM2) decreases from 4.09% to 1.75% when comparing the results of the DeepPixBis and the PAL-RW$_{DeepPixBis}$. This finding indicates that fine-grained partial attack labels are helpful for the improvement of PAD performance under such circumstances. Moreover, PAL-RW$_{MixFaceNet}$ achieves the best overall performance (8.51% ACER value). Note that the only difference between PAL-RW$_{MixFaceNet}$ and PAL-RW$_{DeepPixBis}$ methods is the backbone network architecture. Hence, the lower ACER value obtained by PAL-RW$_{MixFaceNet}$ indicates the efficiency of MixFaceNet [18] and rationalizes our choice of this efficient backbone based on the different sizes of convolutional kernels (thus the different capture levels of attack clues). In addition to comparing the results at a specific operation point, we also present the ROC curves for further observation on a wide range of decision thresholds. Figure 4.12 illustrates the performance of different PAD methods. The pink (PAL-RW$_{DeepPixBis}$) and grey curves (PAL-RW$_{MixFaceNet}$) possess significantly larger areas under the curves than other methods and score lower BPCER values than the baseline methods, especially at low APCER values, which is consistent with the observation of Table 4.7. This work demonstrated that the proposed PAL and RW

Method	Threshold @ BPCER 10% on all data in dev set								
	BPCER (%)		APCER (Print) (%)			APCER (Replay) (%)			
	BM0 54 vids	BM1 108 vids	AM0 486 vids	AM1 972 vids	AM2 162 vids	AM0 972 vids	AM1 1944 vids	AM2 324 vids	ACER (%)
LBP [183]	26.32	11.40	31.38	44.44	36.84	36.74	34.39	28.95	26.33
Inception$_{\text{FT}}$ [213]	1.75	7.02	35.28	30.80	11.70	54.09	52.17	10.23	23.85
CPqD [13]	3.51	7.89	27.49	30.41	16.37	46.20	44.50	10.23	21.75
FASNet$_{\text{FT}}$ [172]	1.75	17.54	10.72	12.77	5.85	30.60	28.09	3.80	16.85
Inception$_{\text{TFS}}$ [213]	8.77	18.42	0.78	1.56	2.34	3.90	5.23	2.63	9.40
FASNet$_{\text{TFS}}$ [172]	14.04	29.82	4.09	3.41	9.36	4.69	2.88	3.80	14.15
DeepPixBis [93]	12.28	31.58	1.75	0.29	4.09	0.88	1.37	0.29	13.13
PAL-RW$_{DeepPixBis}$ (ours)	18.07	28.95	0.00	0.19	1.75	0.00	1.37	0.29	12.99
PAL-RW$_{MixFaceNet}$ (ours)	7.02	4.39	8.58	8.28	10.53	15.98	11.33	17.54	**8.51**

Table 4.7.: The PAD performance of our proposed PAL-RW methods using two network backbones on the CRMA database. The first two columns represent bona fide samples, and the left columns represent different attack types. The APCER and BPCER value is determined by a pre-computed threshold. This threshold is achieved at fixed BPCER 10% on all (masked and unmasked) data in the development set. The number of videos (denoting vids) of each category is noted in the header. The bold number indicates the lowest ACER value. Our PAL-RW$_{MixFaceNet}$ method outperforms other methods.

components do substantially, and on multiple backbone networks, enhance the accuracy of PAD decisions when facing masked faces. We also show that the proposed PAL-RW exhibits better generalization when dealing with masked faces than other established PAD methods. One of the advantages of our proposed method is that PAL-RW is not related to the network structure or training strategy and thus can be easily incorporated into any custom-designed network. Even though the proposed method is well-suited for masked face attacks, the PAL-RW method still has several limitations and can be improved in the future. First, the ground truth for partially masked attacks (AM2) is roughly generated based on 68 face landmarks and is only suitable for attacks with face masks. Therefore, it is worthwhile to produce accurate partial attack annotations (including other types of partial attacks) either manually or specifically designed. The fine-grained ground truth will enhance the generalization ability of PAD to unknown attacks. Second, in our case, the region weight map is set manually for all types of attacks, which is sub-optimal for the final PAD decision. One possible future work is to automatically perform regional weighted inference for different attack types, such as utilizing the position attention map. A partially masked face attack can be considered as an occlusion PAD problem. Although current PAD algorithms have achieved good performance on 2D attacks (print/replay) or 3D mask attacks, the occlusion PAD problem is still understudied. For example, the relevant partial attack data are insufficient. We stress that building a partial attack database is necessary

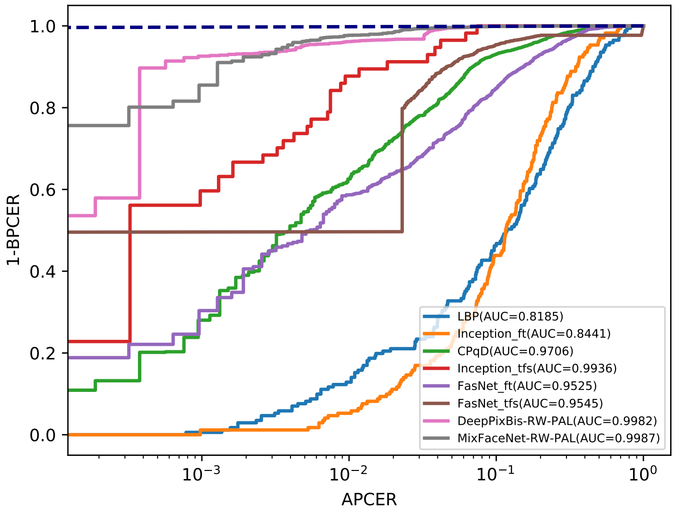


Figure 4.12.: ROCs of different PAD methods and our proposed PAL-RW methods on the CRMA database. PAL-RW$_{DeepPixBis}$ and PAL-RW$_{MixFaceNet}$ scores the lowest BPCER values (especially at low APCER) and achieve the largest areas under curves indicating the best overall performance.

for improving the generalizability of models to unknown attacks.

4.6.4. Discussion

To enable the study, a new large-scale face PAD database, namely CRMA, was presented, including the conventional unmasked attacks, novel attacks with faces wearing masks, and attacks with real masks placed on spoof faces. It consists of 13,113 high-resolution videos and has a large diversity in capture sensors, displays, and capture scales. Moreover, three experimental protocols were designed to study the effect of wearing a mask on the PAD algorithms. With the throughout investigations and analyses, this chapter provided answers to *RQ3*, *RQ4* and *RQ5*.

- This chapter presented a thorough analysis of the vulnerability of FR systems to such novel face masked attacks. The results indicate that FR systems are vulnerable to both masked and unmasked attacks. For example, when the reference images and the computation of system threshold are based on unmasked faces (BM0-BM0), the IAPMR values for unmasked attacks (AM0), masked attacks (AM1), and attacks covered by a real mask (AM2) are 98.40%, 81.60%, and 97.10%, respectively. This leads to the interesting observation that all the investigated FR systems are more vulnerable to attacks where real masks are placed on attacks (AM2) than attacks of masked faces (AM1).

- With the set of extensive experiments conducted in this chapter by using designed protocols, the results indicated that PAD algorithms have a high possibility of detecting masked bona fide samples as attackers (median BPCER value for BM1 in protocol-1 is 48.25%). Moreover, even if the PAD solutions have seen the masked bona fide data, the PAD algorithms still perform worse on masked bona fide samples compared with unmasked bona fides. Furthermore, the PAD solutions trained on masked face attacks (AM1) do not generalize well on attacks covered by a real mask (AM2). For example, the APCER values achieved by DeepPixBis increased from 0.62% for AM1 to 2.92% for AM2 in print attack and from 1.92% for AM1 to 6.43% for AM2 in replay (protocol-3).
- This work presented a solution to target the masked presentation attacks, especially partially covered attacks, by proposing both the partial attack supervision and the regional weighted inference. The goal of partial attack supervision was to guide the neural network to better convergence while training. Meanwhile, regional weight further improved the generalization ability of the model during inference. The detailed ablation study showed the consistent benefits of both components, separately and jointly, in a single solution. Our PAL-RW based models outperformed other established PAD methods when dealing with the possibility of masked faces in PAD decisions.

4.7. Summary

Recent FR studies indicated that the COVID-19 pandemic rendered the conventional FR solutions less effective in many cases. In contrast to the growing number of masked FR studies, the impact of face masked attacks on PAD has not been explored. To target this concern, this chapter studied the vulnerability of FR and the behavior of PAD methods on different types of masked face images. First, to enable the study, a new large-scale face PAD database, CRMA, was presented. The CRMA database includes the conventional unmasked attacks, novel attacks with faces wearing masks, and attacks with real masks placed on spoof faces. Moreover, the CRMA database consists of 13,113 high-resolution videos across diverse capture sensors, displays, and capture scales. Second, extensive experiments were conducted to explore the vulnerability of FR by adopting three conventional FR methods.

The results suggested that FR systems are vulnerable to both masked and unmasked attacks, especially more vulnerable to partially masked attacks than attacks of masked faces, providing an answer to *RQ3*. Third, three protocols were designed to measure the generalizability of the current PAD algorithms on unknown masked bona fide or attack samples, the performance of PAD solutions where the face masks are known, and the generalizability of models trained on masked face attacks when tested on attacks covered by a real mask, respectively. The results obtained by seven PAD methods under three protocols showed that the performance of the existing PAD algorithms, even when involving masked attacks in the training phase, significantly degraded, providing an answer to *RQ4*. In particular, they had a high possibility of detecting masked bona fide samples as attackers. Hence, a PAD method was proposed to target the masked face PAD limitations based on two novel components, the partial attack label supervision, and the regional weighted inference. Conceptually, the proposed solution focused on fine-grained training ground truth information and post-processing predictions, which can be simply incorporated into any common or specially designed neural network architecture. The results demonstrated on the CRMA database indicated improved performance in the perspective of a set of established PAD solutions as a response to *RQ5*. The next chapter will address the generalizability of iris PAD algorithms under the unknown scenarios.

5. Generalizability of PAD algorithms

The previous chapter provided an in-depth analysis of the effects of face masks on the vulnerability of FR and performance of PAD algorithms and presented a solution to target the masked face PAs. This chapter focuses on the analysis and enhancement of generalizability of PAD methods, from face and iris aspects. First, this chapter presents a face PAD solution by intelligently leveraging information from the frequency domain as a response to *RQ6*. Second, this chapter presents two iris PAD solutions as responses to *RQ7* (Section 5.4), and *RQ8* (Section 5.5). This chapter is based on [74, 77, 78, 79, 80, 82, 83].

5.1. Introduction

This chapter is concerned with one face PAD solution and two iris PAD solutions targeting the enhancement of the generalizability of PAD. This section briefly introduces the face and iris-based PAD.

5.1.1. Face presentation attack detection

In recent years, face recognition systems have been widely used in our daily lives for person authentication or access control due to their convenience and remarkable accuracy. However, most existing face recognition systems are vulnerable to PAs. Attackers can use different PAs to impersonate someone or obfuscate their identity. PAs such as print, replay, or 3D mask attacks have been shown to be a serious threat to face recognition systems. Therefore, face PAD plays a critical role in the security of face recognition systems. PAD methods can be broadly categorized into ones based on hand-crafted features [17, 52, 160, 173, 184], and ones based on deep-learning [65, 93, 168, 250, 251]. Hand-crafted based methods utilized traditional texture features such as LBP and its extended versions [17, 173, 184] that are robust to some variations, e.g., color texture, noise artifacts, in PAs. However, the extracted features may not be discriminative enough between bona fide and attacks. Recent PAD studies [65, 168, 245, 251] are competing to boost the performance using CNNs to facilitate more discriminative feature learning. However, CNN-based methods have been a risk of overfitting and thus affect the performance generalization over variations, such as unseen sensors or varied illumination conditions. Considering the characteristics of the hand-crafted and deep learning-based features, it is worth exploring the integration of both features for more discriminative and generalized PAD decisions. In addition to widely used LBP features, several studies [34, 35, 160] attempted to transform images to the frequency domain. Li *et al.*[160] utilized the dissimilarity in Fourier spectra by considering that less high-frequency components exist in attacks compared to bona fide samples. These hand-crafted features are less relevant to the advanced semantic information like identity information but more relevant to the capture conditions, like displayed screen, used photo, or capture sensors. However, most existing hand-crafted features are extracted by static filters, which might limit the representation capacity and make capturing the relevant patterns harder. A recent study [186] proposed the adaptive partition of images in the frequency domain based on a set of learnable frequency filters to detect face forgery clues. In this work, several learnable filters are adopted to capture the PAs cues. Considering the great progress achieved by the deep learning-based methods, we successfully aim at using CNNs to learn subtle differences between bona fide and attacks on both decomposed components in the frequency domain and RGB images in the spatial domain. Recently, attention mechanisms were proposed to model the interdependencies between the channel and spatial features on feature maps of CNNs. Woo *et al.*[236] proposed a CBAM that can be integrated into any CNN architectures and is end-to-end trainable along with the base CNN. The intermediate feature map is adaptively refined by a combination of channel and spatial wise attention. However, most existing attention-based networks do not consider the nature of features in different layers. The features become more abstract and complex when moving from lower to higher layers in a CNN. The features in the lower layers are relevant to the texture information (e.g., edges), and the features in the higher layers emphasize advanced semantic information. Therefore, simply using a combined channel and spatial attention module may be sub-optimal. In this work, different attention modules are successfully applied according to the nature of the deeply learned features.

This work aims to integrate learned features from the frequency and spatial domains for better PAD generalization capability, as a response to *RQ6*. This is achieved through three main aspects. First, a dual-stream PAD solution is proposed based on learnable multi-level frequency decomposition and the proposed hierarchical attention mechanism to capture discriminative and generalize features from both the spatial and frequency domains, namely the LMFD-PAD. Second, an evaluation in both intra-dataset and cross-dataset settings is provided that demonstrates the superiority of the model in cross-dataset PAD when compared to the state-of-the-art, including the PAD methods explicitly targeting domain adaption/shift problem. Third, an ablation study successfully demonstrates the benefits of the proposed LMFD-PAD components, in a step-wise manner, to the cross-dataset PAD performance.

5.1.2. Iris presentation attack detection

Iris recognition systems have been deployed in many law enforcement or civil applications in recent years. However, recognition systems are vulnerable to presentation attacks [48, 176, 243]. A presentation attack is performed to obfuscate the identity of the attacker or impersonate a specific person, such as printed image attacks, textured contact lens attacks, video replaying attacks or synthetic eyes. Despite the high interest from researchers and system vendors in this issue, current algorithms still have some limitations. First, most iris PAD systems only considered iris images captured under a controlled environment [49, 110, 238]. Second, the number of smartphone users worldwide today surpasses three billion and is forecast to grow further by several hundred million in the next few years [112]. Considering the wide usage of mobile devices, algorithms with high computational requirements are challenging to deploy on mobile devices regardless of their high accuracy. Moreover, there are many varying attributes to deal with in the real-world scenario, for example, new types of attacks, different sensors, or capturing conditions, as shown in **??**. Therefore, a good iris PAD system should not only be robust and efficient, but also have generalizability to adapt to changing scenarios.

One challenge in iris PAD is the detection of texture contact lenses, especially when they are confused with transparent lenses. Wearing a cosmetic lens is an easy method to conceal the original texture pattern of the iris and significantly decreases the rate of genuine match rate [6, 142, 189, 243]. That makes developing a robust and accurate presentation attack detection system an essential and valuable task in real-world scenarios. Hughes and Bowyer [121] proposed the first approach to analyze iris images with contact lenses in the 3D space context. Connell *et al.*[157] presented a structured light projection method to detect contact lenses by exploiting the anatomy and geometry of the human eye. Several works [70, 156, 206] investigated the impact of the transparent contact lenses on the iris PAD performance. Their experimental results indicated that a transparent contact lens attack hampered the performance of an iris recognition system. In fact, one can notice small distinguishable artefacts induced by the different types of lenses. With these motivations, Micro Stripes Analyses (MSA) is introduced to target the issue of the iris PAD with textured contact lenses and a lightweight model, and perform a detailed cross-database and unknown attack detection evaluation. The extensive experiments of MSA under various scenarios provide an answer to *RQ7*.

Most of the recent iris PAD solutions trained models by binary supervision, i.e., networks were only informed that an iris image is bona fide or attack, which may lead to overfitting.

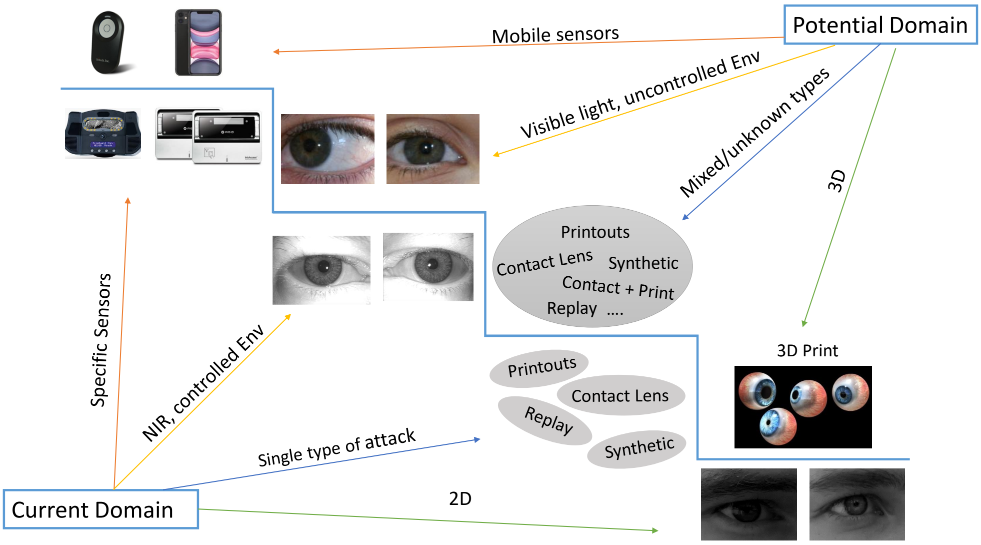


Figure 5.1.: Cross-domain in iris PAD system. We show the currently well-studied domains and the PAD domains where there is open research Potential for novel improvements.

Besides, the limited binary information may be inefficient in locating the regions that contribute the most to making an accurate decision. To target these issues, an attention-based pixel-wise binary supervision (A-PBS) is introduced in Section 5.5. Given that the integration of iris recognition in smart devices is on the rise [146, 190, 201], the study of iris PAD under the visible spectrum is essential. Furthermore, knowing that most iris PAD solutions are developed and trained for images captured in the Near-Infrared light (NIR) domain, an investigation of cross-spectrum iris PAD performance is much needed. To the best of our knowledge, there is no existing work on that investigated the PAD performance under a cross-spectrum scenario. As a result, this work further addressed a visible-light-based iris PAD and the cross-spectrum PAD scenario. The extensive experimental results provide an answer to *RQ8*.

This chapter is organized as follows: Section 5.2 presents a detailed look into related works to face and iris PAD solutions. Section 5.3 introduces a face PAD solution that intelligently leveraging the information from the frequency domain. Section 5.4 introduces a framework for detecting iris presentation attacks that focuses on detecting attack clues around the outer iris boundary based on normalized multiple micro stripes. Section 5.5 presents an attention-based deep pixel-wise binary supervision method (A-PBS) for iris

PAD and provides a comprehensive analysis of generalizability of iris solutions under cross-dataset and cross-spectrum scenarios. A final discussion of the chapter contributions is presented in Section 5.6.

5.2. Related work

This section provides a briefly background information of face PAD and iris PAD separately.

5.2.1. Face PAD

The most relevant prior works are reviewed by focusing on feature-based and deep learning-based face PAD methods, especially those aiming to demonstrate cross-dataset generalizability.

Feature-based methods: Hand-crafted features, such as LBP and image distortion, are utilized broadly to detect presentation attacks. For instance, the commonly used LBP projects the faces to a low-dimension representation and has shown good performance on Idiap Replay-Attack dataset [43]. Boulkenafet *et al.* [13] held an IJCB Mobile Face Anti-Spoofing (IJCB-MFAS) competition [13] carried out on the publicly available OULU-NPU dataset [14] in 2017. The goal of the competition was to evaluate the generalizability of PAD algorithms in a mobile environment. The best performing algorithm among all protocols, named GRADIANT, fused color, texture, and motion information from different color spaces. In addition to LBP, transforming face images into the frequency domain was also previously used. Jourabloo *et al.* [141] used Fast Fourier Transform to analyze the spoofing noise. They found that low-frequency features are related to the color distortion and replay artifacts, while high-frequency responses were more obvious on print attacks. Recently, Chen *et al.* [35] fused the high and low-frequency features for advanced generalizability of face PAD. In their work, three fixed filters were used to extract the high-frequency information from the input images, and low-frequency features were extracted by Gaussian blur filters. However, the hand-crafted and fixed filters might fail to cover the complete frequency domain, and it is hard to use them to capture features adaptively. Thus, Qian *et al.* [186] proposed a set of learnable frequency filters for face forgery detection. In our work, we adapt three learnable filters as suggested in [186] and add one more general filter to obtain the frequency-aware decomposed image components, which is complemented by RGB images.

Deep learning-based methods: Deep learning-based methods have been pushing the frontier of face PAD research and have shown significant improvement in PAD performance. George *et al.*[93] proposed a PAD based on pixel-wise and binary supervised (DeepPixBis) training. However, the DeepPixBis method did not generalize well on unseen attacks/sensors scenarios. To further improve the intra-dataset performance and increase the generalization capability, some studies use auxiliary information, e.g., depth [251] and Remote Photoplethysmography signals [168], for training supervision. For example, Yu *et al.*[250] proposed Neural Architecture Search based method for face PAD (NAS-FAS) based on their previous work on Center Difference Convolution Network (CDCN) [251]. They obtained significantly improved results in both intra-dataset and cross-dataset experimental settings. However, the expensive computation cost of Neural Architecture Search (NAS) must be considered, and the higher error rates in the cross-dataset scenarios suggest that the generalizability is still an open problem. Several methods explicitly targeted the domain generalization problem as an inherent domain shift that can be found between different face PAD datasets. Saha *et al.*[200] proposed a class-conditional domain discriminator module to generate discriminative bona fide and attack features to tackle the domain shift problem. Most domain generalization face PAD methods [159, 200, 203, 204] performed experiments on four publicly available dataset: Oulu-NPU [14], CASIA-MFSD [261], Idiap Replay-Attack [43], and MSU-MFSD [233]. We follow this cross-dataset setting to compare the presented face PAD method against those SOTA methods later in this chapter (as reported in Section 5.3).

5.2.2. Iris PAD

In a recent work, Czajka and Bowyer [48] presented a comprehensive assessment of the SOTA in iris PAD field. They observed that in recent years there had been growing interest in deep learning-based iris PAD. Most neural network-related works achieved superior results than approaches using traditional hand-crafted features, such as LBP [97, 255] and BSIF [142, 176, 188]. We summarize some of the recent iris PAD papers in the last two years (See Table 5.1) and notice that only several works performed cross-database evaluation, as will be explained in details.

The LivDet started in 2013 and the most recent edition took place in 2017 [243]. The LivDet-Iris-2013 and LivDet-Iris-2015 [244] were launched to assess the performance of iris PAD algorithms. Recently, most algorithms achieved high performances on NIR databases. However, by observing databases in LivDet-Iris-2017, it can be noticed that the cross-domain problem, which referred to cross-database and cross-sensor, is considered a major problem for current iris PAD solutions. New databases and novel acquisition conditions of data possessed a higher difficulty than the previous competition. In the

Author	Year	Algorithm	Used Databases			Cross Validation
			Name	Type of Attack	Light Condition	
Chen and Ross [36]	2018	Multi-task CNN	ND-Contact, BERC-Iris-Fake, LivDet-Iris-2017-Clarkson [243], CASIA-Iris-Interval &Syn, LivDet-Iris-2015-Warsaw	PP, CL, SY	NIR	Yes, LE
Hoffmann *et al.* [110]	2018	Multi-patch CNN	LivDet-Iris-2015-Warsaw, CASIA-Iris-Fake, BERC-Iris-Fake	PP, PE, CL	NIR	Yes, LE
Yadav *et al.* [238]	2018	Fusion of VGG features and Multi-level Haralick features	Combined Iris Datbase [238]	PP, CL, SY	NIR	No
Yadav *et al.* [242]	2019	DenseNet based	Combined Iris Database [238], MUIPAD [241], WVU unMIPA [242]	PP, CL	NIR	No
Ferreira *et al.* [89]	2019	wLBP+MLP with adversarial learning	VSIA [188]	PP, RA	VIS	No
Kuehlkam *et al.* [155]	2019	Ensemble multi CNNs fed with mBSIF features	LivDet-Iris-2017 [243]	PP, CL	NIR	Yes, LE
Mandalapu *et al.* [174]	2019	Texture-based Evaluation	LivDet-Iris-2017 [243] without WVU subset	PP, CL	NIR	Yes
McGrath *et al.* [176]	2018	BSIF+ensemble classifiers	NDCLD2015, LivDet-Iris-2017 (Clarkson,IIITD)	CL	NIR	Yes, LE
Czajka *et al.* [49]	2019	Photometric Stereo Features	NDCLD-2015 [142]	CL	NIR	No
Hoffmann *et al.* [111]	2019	Fusion of specific-patch CNNs	LivDet-Iris-2015 Warsaw [244], BERC-Iris-Fake, IrisID	PP, PE, RA, CL	NIR	Yes, LE
Kimura *et al.* [148]	2020	Hyperparameter tuning	LivDet-Iris-2017 [243]	PP, CL	NIR	Yes
Sharma and Ross [205]	2020	DenseNet-based based	LivDet-Iris-2017 [243]	PP, CL	NIR	No

Table 5.1.: Recent neural network-based iris PAD algorithms including information regarding performing cross-database evaluation. Abbreviations: PP - paper printouts, CL - contact lenses, SY - synthetic irises, PE - plastic eyes, RA - replay attack, LE - limited evaluation

LivDet-Iris-2017 competition [243], the Chinese Academy of Sciences (CASIA) proposed the Cascade SpoofNets to detect iris attacks. One SpoofNet was used for detecting printouts attack, and one was for textured contact lens attack. This method achieved the best performance on the Clarkson and Warsaw datasets. The winner of the IIITD-WVU and Notre Dame datasets did not provide detailed information about their algorithms. Recently, Kimura *et al.* [148] proposed fine-tuning the hyperparameters of the SpoofNets to improve the iris PAD performance. Since only the IIITD-WVU dataset was designed for cross-database evaluation, the cross-attack validation was very limited. Therefore, in this section and this paper, we focus on the iris PAD cross-database and cross-attack evaluation.

McGrath *et al.* [176] proposed an open-source PAD solution to distinguish between bona fide and irises with textured contact lenses. They extracted BSIF features and then classified them by an ensemble of classifiers incorporating Support Vector Machine (SVM), random forest, and multi-layer perceptron. They concluded that ensemble classifiers (7 or 8 models) could be robust to cross-dataset validation, as the results were close to the winner of the LivDet-Iris-2017 competition. However, such an analysis of cross-dataset evaluation have limited implications since they only demonstrated experiments by training on one of the databases and evaluating on the other, and not vice versa. Moreover, one of the used databases was the Clarkson database that is no longer publicly available. Furthermore, they neglected the printout attacks from the original datasets and only focused on textured lenses, and thus did not perform cross-attack evaluation. Czajka *et al.* [49] presented the first photometric stereo-based iris PAD method to detect textured contact lenses. They concluded that their proposed algorithm had good generalization capabilities because it was not trained on any specific texture lens pattern. Though they showcased a good performance on regular (dot-like pattern) and irregular (no dot-like patterns) of textured contact lenses, their experiments did not involve any cross-database, cross-sensor, or cross-attack validation. Therefore, the generalizability of the algorithm is insufficiently analysed. A multi-task CNN framework [36] was proposed to estimate iris bounding boxes and detect iris presentation attacks simultaneously. This CNN model was trained on a subset of LivDet-Iris-2015-Warsaw, then tested on the BERC-Iris-Fake and LivDet-Iris-2017-Clarkson database, but not vice versa. Therefore, the demonstration on cross-database capabilities is limited. Similar to the multi-task CNN, Hoffmann *et al.* [110] trained a multi-patch CNN network to learn a mapping relationship between patches of iris pixels and the corresponding classification label. The best performing CNN from the intra-database testing scenario was considered as the trained model and tested against the other two databases. However, the cross-database experimental results showed that the model lost significant performance when tested on the plastic or textured contact lens presentation attacks. Therefore, Hoffmann *et al.* pointed in their work that cross-database PAD is a challenging scenario. In their following work [111], the authors sampled 20 patches, 12 patches from the ocular region and 8 patches from the iris/sclera boundary, to train the multi-patch CNN presented in [110]. They validated the difficulties of cross-domain again in this paper, i.e., the iris PAD algorithms need to account for variations in the sensors, image acquisition environment, subject population, and presentation attack generation procedures. However, extracting small size patches might lose the integrity information of the iris-sclera boundary. In contrast, our stripes ensure the consistency of the captured information, by ensuring a consistency localization. Moreover, the authors reported results on an outdated database (BECR-IF 2006 [158]), no longer publicly available database (Warsaw-2015 [244]), and a private database (IrisID), which limited the possibility of a direct comparison.

In contrast to only using neural networks, several researchers attempted to combine hand-crafted and deep learning features. For example, Yadav *et al.* [238] fused the Multi-level Haralick (MH) texture features with VGG [207] features to encode the textural variations between bona fide and attack iris images. Although this fusion method achieved good performance on each database, the generalization (cross-database) ability of this approach is unclear. Moreover, the VGG-16 [207] network contains 138 million parameters, which indicated that it is hard to deploy on embedded mobile devices, e.g., smartphones. Additionally, a key problem with the above mentioned approaches on iris PAD is that these methods were demonstrated only in a controlled acquisition environment and lacked analyses in varying conditions. As a follow-up work on [238], Yadav *et al.* presented a challenging database [242], which is captured by a mobile iris sensor in varying environmental situations (indoors and outdoors). In addition, they proposed the DensePAD framework [242] to detect iris presentation attacks and demonstrated its efficacy. DensePAD framework tackled the issue of textured contact lens attack, but neglected information beyond the iris boundary. Furthermore, Sharma and Ross [205] exploited the architectural benefits of DenseNet [117] to propose an iris PA detector (D-NetPAD) evaluated on a proprietary database and the LivDet-Iris 2017 databases. With the help of their private additional data, the fine-tuned D-NetPAD achieved good results on LivDet-Iris 2017 databases, however, scratch D-NetPAD failed in the case of cross-database scenarios. Ferreira *et al.* [89] used adversarial learning fed with weighted LBP features to detect iris presentation attacks and demonstrated the algorithm on Visible Spectrum Iris Artefact (VSIA) database [188]. Both works did not discuss the influence of cross-database evaluation. An ensemble of multi-view learning classifiers approach [155] was proposed for cross-domain iris PAD in 2019 and surpassed the winner of LivDet 2017 in all databases. They trained multiple CNNs fed with multi-size BSIF representations and combined those results via meta-fusion. However, the cross-database evaluation using this fusion method did not result in satisfying accuracy. They concluded in the paper [155] that training CNN predictors and testing on another database resulted in accuracy no better than random prediction. Furthermore, Mandalapu *et al.* [174] performed an empirical evaluation of texture-based print and contact lens attacks on the LivDet-Iris-2017 Database. The generalization ability was mainly discussed with different experimental scenarios in this work. However, their preliminary analyses were only based on a database captured under controlled environment, since they did not use the LivDet-Iris-2017 WVU subset. Recently, Chen *et al.* [37] proposed an attention-guided iris PAD method to refine the feature maps of DenseNet [117]. However, this method utilized conventional sample binary supervision and did not report cross-database and cross-spectrum experiments to prove the generalizability of the additional attention module.

Limitations: Based on the recent iris PAD literature, it can be noticed that there is a small error performance gap to be closed in NIR based iris PAD. Moreover, the uncontrolled iris captures are still a challenge for iris PAD, especially with limited computational resources. Furthermore, the deep-learning-based methods boost the performance but still have the risk of overfitting under cross-PA and cross-database scenarios. One of the major reasons causing overfitting is the lack of availability of a sufficient amount of variant iris data for training networks. Another possible reason might be binary supervision. While the binary classification model provides useful global information, its ability to capture subtle differences in attacking iris samples may be weakened, and thus the deep features might be less discriminative. In addition to the challenges across database scenarios, another issue is that there is no existing research dedicated to exploring the generalizability of PAD methods across spectral scenarios. PAD research in the NIR domain [77, 155, 205, 238] has attracted much attention, while few studies [190, 240] investigated PAD performance in the visible spectrum. Furthermore, the generalizability of PAD methods under the cross-spectrum is unclear. Suppose a model trained on NIR data can be well generalized to visible-light data. In that case, it requires only low effort to transfer such solution to low-cost mobile devices [146, 190, 201], which simplifies its application in the real world. In addition to the model generalizability problem, another concern is the unclear fairness in PAD systems due to the insufficient annotated and balanced data resources. Despite the limited database availability, the demographic bias in iris PAD is an open issue and is of great interest. With the widespread use of iris recognition technology, we believe the bias problem needs to be analyzed to enable future mitigation efforts.

The following sections introduce first a face PAD solution that intelligently leverages information from the frequency domain to enhance the generalizability of PAD. Then, two iris PAD methods are presented to target the generalizability of iris PAD under cross-domain scenarios.

5.3. Enhancing the generalizability of face PAD

This section presents a learnable multi-level frequency decomposition based face PAD method, LMFD-PAD, targeting the generalizability of PAD performance. A dual-stream network architecture was employed. The first stream learned discriminative features in the frequency domain by using learnable frequency filters to obtain frequency decomposed image components, while the other stream used RGB face images as input to learn features in the spatial domain. Moreover, the hierarchical attention mechanism was presented to fuse features from both domains at different stages of the network. A spatial attention module was added at the lower layers of the CNN to capture the texture features, and the channel attention module was added at the higher layers of CNN to obtain advanced semantic information. The experiments were demonstrated under intra-dataset and cross-dataset settings. The presented LMFD-PAD method achieved comparable results in intra-dataset scenarios. Moreover, in most cross-dataset cases, the proposed solution outperformed SOTA face PAD methods, including the methods addressing the domain adaption/shift and generalization capability problem. The proposed components of the LMFD-PAD solution were additionally proved in a step-wise ablation study. This section first presented the detailed architecture of LMFD-PAD and the experiment setup. Then, the obtained results will be discussed to respond to *RQ6*.

5.3.1. Methodology

In this section, we will provide details of the LMFD-PAD solution. We will introduce the Multi-level Frequency Decomposition (MFD), including four learnable frequency filters. Then the dual-stream network architecture is introduced where using a hierarchical attention mechanism to integrate the features learned in frequency and spatial domain, and at last present the used loss functions.

Multi-level Frequency Decomposition

Deep-learning based face PAD methods achieved great progress in intra-dataset evaluations. However, the performance normally drops drastically when testing on unseen datasets [185]. This might be caused by the variations in the attacks and capture environments, such as illuminations and sensors. To address this issue, our proposed LMFD solution decomposes an input face image into different level frequency components. Frequency domain analysis is a classical method in image signal processing and has been widely used for general image classification and texture classification tasks [104, 212]. Moreover, some face PAD methods attempted [34, 35, 160] to transform the images in frequency

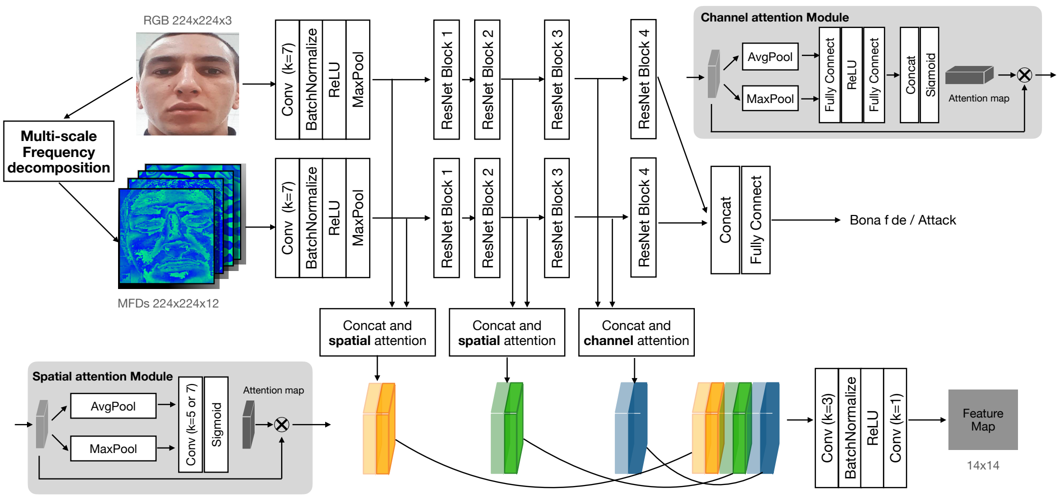


Figure 5.2.: The overall workflow of our proposed LMFD-PAD solution. Note the utilization of our MFD and Hierarchical Attention Mechanism (HAM) (three different channel attention components) components.

domain and mine the artifacts cues. The results showed that features in the frequency domain are less sensitive to the variations of the capture environments (e.g., sensors or light conditions). However, most existing frequency-based face PAD methods used filters with fixed weight and maybe sub-optimal for discriminative feature learning.

In our work, we use a set of adaptively learnable frequency filters described in [186] for face forgery detection. First, N manually designed binary base filters $\mathcal{F}_b = \{f_b^i | 1 \leq i \leq N\}$ partition the frequency domain into low, middle, high frequency bands. The goal of the binary base filters is a roughly equal division of spectrum intp N bands from low frequency to high frequency. Then, N learnable filters $\mathcal{F}_l = \{f_l^i | 1 \leq i \leq N\}$ are added to such binary base filters. The benefit of such learnable filters is the adaptive selection of the frequency of interest beyond the fixed base filters. Finally, a decomposed image component C_i of an input image x can be computed following the equation:

$$\mathrm{C}_i = \mathcal{D}^{-1}\{\mathcal{D}(x) \odot [f_b^i + \sigma(f_l^i)]\}, i = \{1, ..., N\}, \tag{5.1}$$

where $\mathcal{D}$ is DCT, $\mathcal{D}^{-1}$ is inversed DCT, and $\odot$ is the element-wise product. The $\sigma(f) = \frac{1-exp(-f)}{1+exp(-f)}$ is used to normalize the value of f between -1 and $+1$.

In our case, N is set to 4 to obtain explicitly divided frequency domain of low, middle, and high-frequency bands and the complementary full frequency band. Three bands are chosen as described in [186]: 1) the low frequency band f_{base}^1 is the first 1/16 of the entire spectrum, 2) the middle frequency band f_{base}^2 is between 1/16 and 1/8 of the entire spectrum, 3) the high frequency band f_{base}^3 is between 1/8 and 7/8 of the entire spectrum. However, the partitioned frequencies may not be sufficient to obtain subtle cues between bona fide and attacks. Therefore, we add one additional learnable filter f_{base}^4 where the frequency band is the entire spectrum. Moreover, we also keep the input RGB image to provide more visual information and complementary to frequency domain information (as shown in Figure 5.2).

In the experiments, face detection is firstly performed on the input image by MTCNN framework [256]. Then, the detected RGB face image is resized to $224 \times 224 \times 3$ pixels. According to the Equation 5.1, four obtained components are stacked along the channel axis, i.e, the size of a stacked decomposition is $224 \times 224 \times 12$. Then, we utilize dual-stream (RGB and MFD) networks to extract different features in a face image (see Figure 5.2). In our work, we use the ResNet-50 [107] as our backbone network.

Hierarchical Attention Mechanism

So far, we use the dual-stream to learn discriminate features in parallel, which may be sub-optimal for a final PAD decision. To enhance that, we propose our HAM to integrate features from the frequency domain and semantic image domain and to utilize the features from different layers in the dual-stream.

This HAM is inspired by Convolutional Block Attention Mechanism (CBAM) [236], which proposed channel and spatial attention blocks for the general computer vision task, and Attention Pixel-wise Binary Supervision (A-PBS) method [79], which employed and fused spatial attention features from multi-layers for the iris PAD task. The CBAM [236] consisting of the channel, and distinctive spatial sub-modules can be added into networks according to the custom design needs and showed improvements in classification and detection performance with various neural architectures. A-PBS method [79] adopted only spatial attention module (i.e., no channel attention module) aiming to locate the most informative region in an RGB eye image, where might contribute most to a PAD decision. However, in our MFD stream, we have multi-level frequency features, and the weights of filters are adaptively learning while the model is training. The high-frequency component emphasizes features like edges and texture information, while the low-frequency component is related to the spatial distribution of the color gamut. Therefore, channel attention is additionally applied in our framework.

Figure 5.2 shows that spatial attention modules are inserted after the first convolution block and the second ResNet block, respectively, while a channel attention module is added following the third ResNet block. The reason for such attention modules arrangement is based on the nature of the features extracted from different layers. The features from lower to higher layers become more abstract and complex. More specifically, the features in the lower layers are related to the appearance and texture cues, and the features in the higher layers might reveal the semantic content information. Consequently, we perform a spatial attention module on a fused feature in lower layers to focus on texture details like the edge. Then, a channel attention module is added after the third ResNet block to learn the advanced semantic features. To be consistent with the observation on the nature of features in different layers, the size of the convolutional kernel is 7×7 in the first spatial attention module and 5×5 in the second spatial attention module, as the smaller convolutional kernel is more suitable for locating the small-scale texture cues. Finally, the attentive features are fused to preserve richer patterns. Moreover, we use pixel-wise and binary supervision to train the dual-stream networks as suggested in [93] where the intermediate feature map can be considered as the scores generated from the patches in an image and thus improve the performance. On the one hand, the attentive feature maps from different layers are concatenated and fed to the stacked two convolution layers to output a feature map. The size of the output feature map in our case is 14×14 for pixel-wise supervision. On the other hand, the features from the last ResNet block in two streams are also concatenated and fed to the fully connected layer for binary supervision.

Loss function

BCE loss has proved to perform well when used for pixel-wise and binary supervision [93]. Nevertheless, to reduce the sensitivity to outliers in the output feature map, we use the Smooth L1 (SL) function to compute the loss between the output feature map and the ground truth binary mask. For binary supervision, we use the Focal Loss instead of BCE loss because the Focal loss (FL) with a relaxing factor can down-weight easy samples (i.e., samples correctly classified with high confidence) and make the model focus on the hard samples with low classification confidence. The equation for Smooth L1 is shown below as:

$$\mathcal{L}_{SL} = \frac{1}{n} \sum z$$

$$\text{where} \quad z = \begin{cases} \frac{1}{2} \cdot (y-x)^2, & \textit{if} \quad |y-x| < 1 \\ |y-x| - \frac{1}{2}, & \textit{otherwise} \end{cases}$$

where n is the number of pixels in the output map (14 in our case). x and y refer to the values in the output feature map and the ground truth label, respectively. The equation of

Focal loss is:

$$\mathcal{L}_{FL} = -(1-p_t)^{\gamma}\log(p_t)$$

$$\text{where} \quad p_t = \begin{cases} p, & \textit{if} \quad y = 1 \\ 1-p, & \textit{otherwise} \end{cases}$$

where p is the predicted probability when the ground truth label y is 1 (bona fide in our case) and γ is a tunable focusing parameter (γ is 2 in our experiments). The overall loss function is given as:

$$\mathcal{L}_{overall} = \lambda_1 \cdot \mathcal{L}_{SL} + \lambda_2 \cdot \mathcal{L}_{FL} \tag{5.2}$$

For exploring the effect of loss functions, we also report the results of BCE loss as used in [93] as an ablation study (as shown in Table 5.4).

5.3.2. Experimental setup

Datasets:

Our method is evaluated on four publicly available face PAD datasets: Oulu-NPU [14], CASIA-MFSD [261], Idiap Replay-Attack [43], and MSU-MFSD [233] under different scenarios. Oulu-NPU [14] dataset consists of 55 subjects and 5940 videos recorded by six mobile phones. Four protocols are provided to evaluate the generalizability of algorithms. Protocol-1 studies the impact of illumination variations, while Protocol-2 evaluates different attacks created by various instruments. Protocol-3 examines the effect of different capture cameras, and Protocol-4 explores all the challenges above by leave-one-out cross-validation. CASIA-MFSD [261] includes 50 subjects and 600 videos captured by three different quality cameras. This dataset contains three attack types: warped photo attack, cut photo attack, and video replay attack. Idiap Replay-Attack [43] contains 50 subjects and 300 videos captured by different sensors and different illumination conditions. Moreover, two types of attacks are included in this dataset: print and replay attacks. MSU-MFSD [233] contains 35 subjects and 440 videos captured by two different resolutions of cameras. This dataset also includes two types of attacks, printed photo attacks and replay attacks. The videos in datasets are recorded under different environments with variant cameras and subjects, suitable for cross-dataset domain generalization protocol. Moreover, the subjects in the training set and test set are disjoint in intra-dataset settings.

Implementation details:

The proposed dual-stream networks are based on ResNet-50 [107] with pre-trained weights on the ImageNet dataset [67]. The data in all PAD datasets are videos, thus, we sample 10 frames in the average time interval of each video to train and test our method. For each frame, the face is detected and cropped by the MTCNN method [256] and resized to $224 \times 224 \times 3$ pixels. In the training phase, the SGD optimizer is used with an initial learning rate of 0.001, the momentum of 0.9, and a weight decay of 0.0001. Then, the exponential learning rate scheduler is used with a multiplicative factor of the learning rate decay value (γ) of 0.995. The ratio of bona fide and attack data is close to 1:1 by simply duplicating the needed images to reduce the effect of biased data. Several data augmentation techniques are used for better generalization ability, including horizontal flip, rotation, cutout, RGB channel shift, and color jitter. To further reduce overfitting, the early stopping technique is utilized with the maximum epochs of 100 and the patience epochs of 15. The batch size in the training phase is 32. In our experiments, the λ_1 in overall loss function 5.2 is set manually to 1 at the beginning of the training and changed to 100 after five training epochs, while λ_2 is set to 1 in the whole training phase. In the testing phase, a final PAD decision score of a video is a fused score (mean-rule fusion) of all frames.

Evaluation metrics:

We follow the sub-protocols and metrics as defined in the competition [13] which was performed on the OULU-NPU [14] dataset for a fair comparison. The Attack Presentation Classification Error Rate (APCER) [127] is computed separately for each presentation attack instrument (PAI), e.g., print or replay following the equation:

$$ACPER_{PAI} = \frac{1}{N_{PAI}} \sum_{i=1}^{N_{PAI}} (1 - p_i) \tag{5.3}$$

where N_{PAI} is the number of attack samples for a given PAI, p_i is the predicted binary label of the i^{th} presentation (0 for bona fide and 1 for attack). Then, following the OULU-NPU protocol [14], APCER$_{wc}$ is the highest APCER is selected to report the overall performance, i.e., the worst case among all the presentation instruments. The equation is APCER$_{wc}$ = $\max$ (APCER$_{PAI}$) among all PAIs. Bona Fide Presentation Classification Error Rate (BPCER) [127] is the proportion of incorrectly classified bona fide samples. Average Classification Error Rate (ACER) is the mean of APCER$_{wc}$ and BPCER. Moreover, to report the cross-dataset results and to be consistent with previous works [17, 159, 168, 204, 250], we report Half Total Error Rate (HTER) and Area Under the receiver operating Curve (AUC) are used for the cross-dataset domain generalization protocol on OULU-NPU [14] , CASIA-MFSD [261], Idiap Replay-Attack [43] and MSU-MFSD [233] datasets. The HTER is half of the sum of the APCER and BPCER.

Prot.	Method	APCER$_{wc}$(%)	BPCER(%)	ACER(%)
1	GRADIANT [13]	1.3	12.5	6.9
	Auxiliary [168]	1.6	1.6	1.6
	FAS-TD [231]	2.5	0.0	1.3
	STASN [245]	1.2	2.5	1.9
	DeepPixBis [93]	0.8	0.0	0.4
	CDCN++ [251]	0.4	0.0	**0.2**
	SSR-FCN [65]	1.5	7.7	4.6
	NAS-FAS [250]	0.4	0.0	**0.2**
	LMFD-PAD (ours)	1.4	1.6	1.5
2	GRADIANT [13]	3.1	1.9	2.5
	Auxiliary [168]	2.7	2.7	2.7
	FAS-TD [231]	1.7	2.0	1.9
	STASN [245]	4.2	0.3	2.2
	DeepPixBis [93]	11.4	0.6	6.0
	CDCN++ [251]	1.8	0.8	1.3
	SSR-FCN [65]	3.1	3.7	3.4
	NAS-FAS [250]	1.5	0.8	**1.2**
	LMFD-PAD (ours)	3.1	0.8	2.0
3	GRADIANT [13]	2.6 ± 3.9	5.0 ± 5.3	3.8 ± 2.4
	Auxiliary [168]	2.7 ± 1.3	3.1 ± 1.7	2.9 ± 1.5
	FAS-TD [231]	5.9 ± 1.9	5.9 ± 3.0	5.9 ± 1.0
	STASN [245]	4.7 ± 3.9	0.9 ± 1.2	2.8 ± 1.6
	DeepPixBis [93]	11.7 ± 19.6	10.6 ± 14.1	11.1 ± 9.4
	CDCN++ [251]	1.7 ± 1.5	2.0 ± 1.2	1.8 ± 0.7
	SSR-FCN [65]	2.9 ± 2.1	2.7 ± 3.2	2.8 ± 2.2
	NAS-FAS [250]	2.1 ± 1.3	1.4 ± 1.1	**1.7 ± 0.6**
	LMFD-PAD (ours)	3.5 ± 3.2	3.3 ± 3.2	3.4 ± 3.1
4	GRADIANT [13]	5.0 ± 4.5	15.0 ± 7.1	10.0 ± 5.0
	Auxiliary [168]	9.3 ± 5.6	10.4 ± 6.0	9.5 ± 6.0
	FAS-TD [231]	14.2 ± 8.7	4.2 ± 3.8	9.2 ± 3.4
	STASN [245]	6.7 ± 10.6	8.3 ± 8.4	7.5 ± 4.7
	DeepPixBis [93]	36.7 ± 29.7	13.3 ± 14.1	25.0 ± 12.7
	CDCN++ [251]	4.2 ± 3.4	5.8 ± 4.9	5.0 ± 2.9
	SSR-FCN [65]	8.3 ± 6.8	13.3 ± 8.7	10.8 ± 5.1
	NAS-FAS [250]	4.2 ± 5.3	1.7 ± 2.6	**2.9 ± 2.8**
	LMFD-PAD (ours)	4.5 ± 5.3	2.5 ± 4.1	3.3 ± 3.1

Table 5.2.: The results of the intra-dataset evaluation under the four protocols of the OULU-NPU dataset [14]. The bold numbers refer to the lowest ACER in each protocol. Note that our LMFD-PAD achieves competitive performance overall and performs better than most methods that do not use auxiliary information (depth or rPPG).

5.3.3. Results

Comparison with the SOTA methods

Intra-dataset results on OULU-NPU An IJCB-MFAS competition [13] was carried out on the publicly available OULU-NPU dataset. To assess the generalizability of the face PAD methods, four protocols are provided consisting of cross-environment, cross-PAIs, cross-sensors, cross-all scenarios. For a fair comparison, we strictly follow the definition and evaluation metric of those protocols. In this study, we compare our LMFD-PAD method with the best performing method in IJCB-MFAS competition [13], GRADIANT. Moreover, we also compare with several recently PAD methods: Auxiliary [168], FAS-TD [231], STASN [245], DeepPixBis [93], CDCN++ [251], SSR-FCN [65], NAS-FAS [250] proposed from 2018 to 2021. The results are reported in Table 5.2. [1] The LMFD-PAD achieved ACER values of each protocol are 1.5%, 2.0%, 3.4%, and 3.3%, respectively. It can be observed that our method obtain competitive results in comparison to state-of-the-art methods. For example, the lowest ACER in the most challenging Protocol-4 is 2.9% achieved by NAS-FAS [250], while our LMFD-PAD ACER value is 3.3%. This result indicates that our model generalizes well on the cross-test scenarios. Considering that we employ pixel-wise supervision, we can group those PAD methods into three groups based on supervision manner for further comparison. GRADIANT [13] and STASN [245] was trained only by binary supervision. DeepPixBis [93], SSR-FCN [65] and our method utilized the pixel-wise and binary supervision. The left four PAD approaches used depth or/and rPPG supervision. It can be found in Table 5.2 that our method possesses improved performance compared to pixel-wise and binary supervised models in most cases but scored below the depth/rPPG supervised networks in some cases. This might drive an extension of our work by generating depth or/and rPPG information to improve the intra-dataset performance. In this case, however, the trade-off between computational resource/time and performance needs to be considered.

Cross-dataset results In the cross-dataset scenario, four publicly available face PAD datasets: Oulu-NPU [14] (O for short), CASIA-MFSD [261] (C for short), Idiap Replay-Attack [43] (I for shot), and MSU-MFSD [233] (M for short) are used. Three datasets are randomly selected for training and the remained one is used for testing. Specifically, following previous works targeting the domain adaption and generalization capability of face PAD [159, 200, 203, 204], four settings are performed: O&C&I → M, O&M&I → C, O&C&M → I and I&C&M → O.

[1] The results of state-of-the-art solutions listed in Table 5.2 and 5.3 are those reported in their paper.

Method	O&C&I → M		O&M&I → C		O&C&M → I		I&C&M → O	
	HTER(%)	AUC(%)	HTER(%)	AUC(%)	HTER(%)	AUC(%)	HTER(%)	AUC(%)
MS LBP [173]	29.76	78.50	54.28	44.98	50.30	51.64	50.29	49.31
Binary CNN [237]	29.25	82.87	34.88	71.94	34.47	65.88	29.61	77.54
IDA [234]	66.67	27.86	55.17	39.05	28.35	78.25	54.20	44.59
Color Texture [17]	28.09	78.47	30.58	76.89	40.40	62.78	63.59	32.71
LBPTOP [184]	36.90	70.80	42.60	61.05	49.45	49.54	53.15	44.09
Auxiliary(Depth Only) [168]	22.72	85.88	33.52	73.15	29.14	71.69	30.17	77.61
Auxiliary(All) [168]	-	-	28.40	-	27.60	-	-	-
NAS-FAS [250]	16.85	90.42	15.21	92.64	**11.63**	**96.98**	13.16	94.18
MMD-AAE [159]	27.08	83.19	44.59	58.29	31.58	75.18	40.98	63.08
MADDG [203]	17.69	88.06	24.50	84.51	22.19	84.99	27.98	80.02
RFMetaFAS [204]	13.89	93.98	20.27	88.16	17.30	90.48	16.45	91.16
CCDD [200]	15.42	91.13	17.42	90.12	15.87	91.72	14.72	93.08
LMFD-PAD (ours)	**10.48**	**94.55**	**12.50**	**94.17**	18.49	84.72	**12.41**	**94.95**

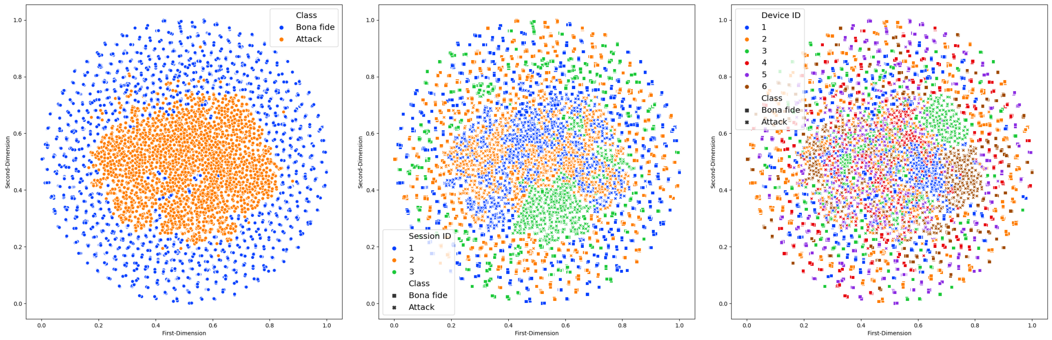
Table 5.3.: The results of the cross-dataset evaluation under different experimental settings on four face PAD datasets. In each setting, three datasets are used for training, and one remaining dataset is used for testing. Our LMFD-PAD method is compared with state-of-the-art face PAD methods reporting on this protocol. Not that four of the state-of-the-art methods MMD-AAE, MADDG, RFMetaFAS, and CCDD are explicitly designed to target the domain shift problem. The bold numbers indicate the lowest HTER and highest AUC in each setting.

In our work, we compare our LMFD-PAD model against eight state-of-the-art face PAD methods including depth/rPPG supervision based Auxiliary [168] and NAS-FAS [250] which outperformed in intra-testing on OULU-NPU dataset [14]. In addition, we also compare our method with four state-of-the-art domain generalization face PAD methods: MMD-AAE [159], MADDG [203], RFMetaFAS [204], and CCDD [200], which explicitly target the domain shift problem. The results are reported in Table 5.3 where the last four methods are face methods addressing domain shift problems. Our proposed LMFD-PAD method achieves significantly improved performance in three experiment settings. For example, the HTER value of our model is 10.48% in O&C&I → M setting and 12.50% in O&M&I → C and 12.41% in I&C&M → O, while the second-ranking results in those settings are 13.89%, 15.21%, and 13.16%, respectively. Although our LMFD-PAD method is not explicitly designed for the domain shift problem, our method obtains better performance than domain generalization face PAD methods in most cases. The cross-dataset results are consistent with the result in the most challenging intra-dataset Protocol-4 of OULU-NPU dataset [14]. We conclude that our method is able to learn more generalized features, which perform well on unseen domains. However, it is still unclear which part of our model benefits the improved results. This question will be answered in the following

RGB	MFD	HAM	BCE	FL+SL	O&C&I → M		O&M&I → C		O&C&M → I		I&C&M → O	
					HTER(%)	AUC(%)	HTER(%)	AUC(%)	HTER(%)	AUC(%)	HTER(%)	AUC(%)
√			√		17.14	90.47	22.12	82.10	24.62	82.28	19.47	88.16
√	√		√		15.47	93.17	17.21	87.50	23.51	83.25	17.26	90.41
√	√	√	√		11.19	93.39	16.83	90.62	21.42	83.92	22.27	85.98
√	√	√		√	**10.48**	**94.55**	**12.50**	**94.17**	**18.49**	**84.72**	**12.41**	**94.95**

Table 5.4.: The results of the ablation study on model inputs, components, and loss functions. The ablation study is performed on cross-dataset experimental settings to uncover the components generalizability benefits. One can note that in most experiments, each of the proposed components contributes positively to the cross-dataset PAD performance.

section by exploring the effect of the MFD, HAM parts, and loss function in an ablation study.



(a) BF and attack.

(b) Varied scenarios.

(c) Varied devices.

Figure 5.3.: t-SNE visualization of a cross-dataset setting I&C&M - O using our LMFD-PAD embeddings, where the test set is OULU-NPU dataset consisting of three capture environments with different illumination conditions and six mobile devices. The first t-SNE plot represents the two classes: bona fide (blue) and attack (orange). The second and third t-SNE plot indicates three capture scenarios and six capture devices, respectively. In Figures 5.3b and 5.3c, each color corresponds to an environment or device, the signs ■ and x refers to bona fide and attack, respectively. It is noted that the embeddings from the LMFD-PAD still find a common pattern between the attacks captured under different settings.

Ablation study on model components

So far, the results in Table 5.2 and Table 5.3 are obtained by our *full model* including the MFD, HAM and a combined loss function of Focal loss and Smooth L1 loss (Equation 5.2). However, the detailed effect of each part is unknown. Therefore, we present an ablation study on model components, and the results are summarized in Table 5.4. This aims at understanding the generalization benefits of each of the proposed components. The training hyper-parameters are the same for all combinations in Table 5.4 (training details are described in Section 5.3.2). Since we assume that our method is able to learn discriminative and generalize features, the ablation study is demonstrated under the cross-dataset experimental setups on four datasets.

Impact of MFD: To explore the effect of the learnable frequency decomposition, we train a one-stream network using only RGB face images as input and a dual-stream network consisting of RGB and MFD, both solutions are trained by minimizing the BCE loss. The results in Table 5.4 shows the improvement by the additional MFD component (the HTER is decreased from 17.14% to 15.47% in the O&C&I → M setting). A consistent performance enhancement is seen under all the experimental setups in Table 5.4.

Impact of HAM: In contrast to learning the features in the image and the frequency domains in parallel and fusing such features just before the classification layer, we add the HAM component to fuse such features earlier followed by different attention blocks according to the levels of layers. The corresponding results are reported in the second row and third row of Table 5.4 where it is noticeable that the addition of the proposed HAM did enhance the performance in most experimental settings.

Impact of loss function: In our LMFD-PAD solution, we use the Focal loss to supervise the binary label prediction and the Smooth L1 to supervise the feature map label prediction instead of the commonly used BCE loss. To explore the effect of such modification, we compare it to using the BCE loss for pixel-wise and binary supervision. The weights of both BCE losses is set to 0.5 as used in [93]. As presented in Table 5.4, the loss combination used in our LMFD-PAD solution strongly enhances the PAD performance across all the cross-dataset experimental settings.

We conclude that our LMFD-PAD full model boosts the performance generalizability further by adding each of the MFD, HAM, and a combined loss function.

Visualization and analysis

In our assumption, the MFD module is able to learn rich generalizable features that adapt well to unseen datasets, especially for unseen sensors or illumination. To further verify this assumption, we use t-SNE [222] plots to visualize deep features in the cross-dataset case I&C&M → O. This setting is chosen because the unseen test set is OULU-NPU dataset [14] consisting of more variation of environment and capture devices and thus it is better for visualization. The deep features are extracted from the last convolution layer before the classification layer, and then the Principal Component Analysis (PCA) is used to reduce the dimensionality of features to 128-D to reduce the computational cost of the t-SNE. Such features are then projected to 2-D features by t-SNE. Figure 5.3 depicts t-SNE plots on two classes (bona fide and attack), three capture environments, and six capture devices from left to right. As seen in Figure 5.3a, bona fides and attacks can be considered as coarsely non-linearly separable. This indicates that our model learns discriminative and generalizes features between bona fides and attacks. In Figure 5.3b, blue, orange, and green represent three environments of various illuminations. It can be seen that different environments are more obviously clustered in the attack category, while they are clustered more randomly in the bona fide category. A similar observation can be found on different mobile devices in Figure 5.3c. These findings suggest that our model is able to mine the general attack artifacts patterns across data capture variety and thus generalizability on unseen datasets is less effect by different sensors or illuminations. This confirms the achieved cross-dataset results reported above.

5.3.4. Discussion

This section proposed a learnable multi-level frequency decomposition based face PAD method, LMFD-PAD, targeting the generalizability of PAD performance. A dual-stream network architecture was employed . The first stream learned discriminative features in the frequency domain by using learnable frequency filters to obtain frequency decomposed image components, while the other stream used RGB face images as input to learn features in the spatial domain. Moreover, the hierarchical attention mechanism was proposed to fuse features from both domains at different stages of the network. A spatial attention module was added at the lower layers of the CNN to capture the texture features, and the channel attention module was added at the higher layers of CNN to obtain advanced semantic information. The experiments were demonstrated under intra-dataset and cross-dataset settings. The presented LMFD-PAD method achieved comparable results in intra-dataset scenarios. Moreover, in most cross-dataset cases, the proposed solution outperformed SOTA face PAD methods, including the methods addressing the domain adaption/shift and generalization capability problem. The proposed components of the LMFD-PAD solution were additionally proved in a step-wise ablation study.

5.4. Micro stripes analyses for iris PAD

This section introduces a framework for detecting iris presentation attacks that focuses on detecting attack clues around the outer iris boundary based on normalized multiple micro stripes [78, 82]. The classification decision is made by the majority vote of these micro-stripes. An in-depth experimental evaluation of this framework reveals a superior performance in three databases, in comparison with SOTA algorithms and baselines. Moreover, MSA solution minimizes the confusion between textured (attack) and transparent (bona fide) lenses presentations in comparison to SOTA methods. We support the rationalization of the proposed method by studying the significance of different pupil-centered eye areas in iris PAD decisions under different experimental settings. In addition, extensive cross-database and cross-attack (unknown attack) detection evaluation experiments are demonstrated to explore the generalizability of our proposed method, texture-based method, and neural network-based methods in three different databases. The results indicate that the presented MSA approach has, in most experiments, better generalizability compared to other baselines, as a response to *RQ7*.

5.4.1. Methodology

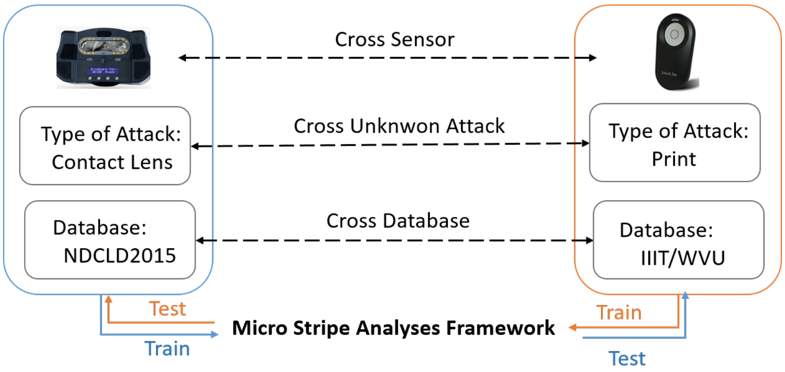


Figure 5.4.: Cross-database and unknwon attack detection scenarios.

There are many methods proposed to address the iris PAD problem in the literature. However, only a few works have been investigated the generalizability of iris PAD approaches from aspects such as the type of presentation attacks, sensors, databases. Although most algorithms achieved high accuracy on databases captured in the controlled environment with NIR illumination, the performance on the different databases has a significant difference. Compared to intra-database evaluation (i.e., training on one database and testing a subject-disjoint set of the same database), cross-database evaluation (i.e., training on one or more databases and testing on other databases.) is considered as a more difficult challenge. Firstly, because of the limitation of databases, trained models are easily overfitting even with regularization techniques. Moreover, many variations need to be considered in the iris PAD field, such as characters of sensors, data acquisition environments, subject biases, and unseen presentation attacks. Consequently, algorithms with high generalizability and transferability are an urgent requirement in real-world scenarios. These algorithms should have the ability to adapt to previously unknown data and maintain similar results. As mentioned in the related works, cross-database validation has gradually raised some attention. Nonetheless, most previous works [36, 110, 111, 155, 176] did not perform extensive experimental analyses. Usually, the best-trained model in one training database was chosen and tested in another one. Furthermore, a recent paper [174] has investigated the generalizability of the textured-based iris PAD methods in cross-database and cross-attack settings. Nevertheless, to our knowledge, there is no extensive deep-learning-based cross-database evaluation so far for iris PAD approaches.

Hence, as shown in Figure 5.4, an extensive experimental analyses for deep learning-based iris PAD methods with cross-database, cross-sensor and cross-attack setting is demonstrated in this work. In order to examine the generalizability between hand-crafted features and neural network features based approaches, four iris PAD methods and three databases are used. Three databases consisting of bona fide and three different kinds of presentation attacks (contact lens, printouts, contact lens + printouts) from various sensors experiment in 3 different scenarios. In the following subsections, we discuss our proposed MSA solution, one hand-crafted feature-based baseline, and two neural network-based methods.

Proposed MSA method

In this subsection, we describe our MSA framework by focusing on how overlapping micro-stripes can be used to address the iris PAD problem. Figure 5.5 presents the overall framework for our MSA solution. It starts with a captured iris image, followed by a number of preprocessing steps. These steps include coarse iris segmentation, segmentation extension, iris normalization, and overlapping micro-stripes extraction. The processed micro-stripes are fed to a specifically trained neural network, which results are fused in a majority vote process. The following subsections present these steps in more details.

Iris Segmentation and Normalization Iris images do not only contain the iris region but also exhibit additional information around the iris, e.g., contact lens edges. Previous

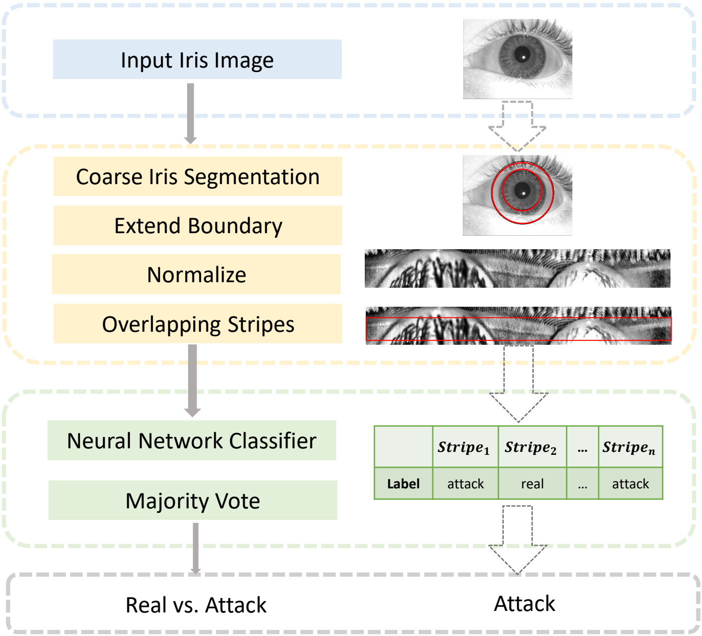
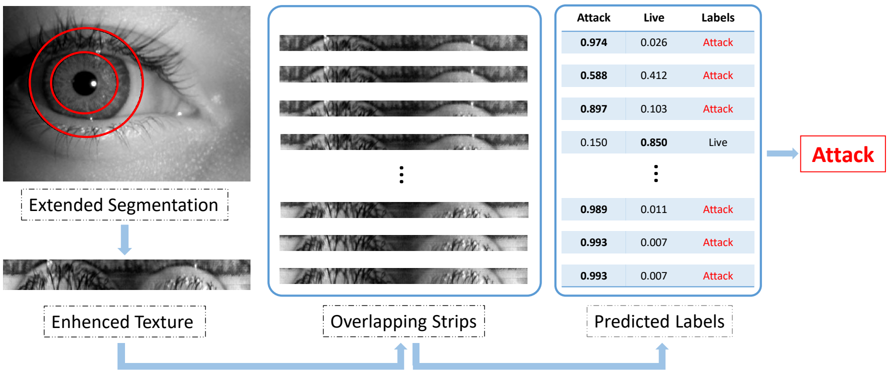


Figure 5.5.: Architecture of the proposed MSA algorithm for iris PAD. Input iris image is from the NDCLD-2015 database [142].

works noticed this property and offered a solution aiming to detect contact lens reflective properties [116, 157]. Moreover, to further motivate our MSA normalization process, Hoffmann *et al.* [111] found that the iris/sclera boundary comprised the presentation attack artifacts, such as the clear white, unobstructed boundary, in many plastic eyes. Typically, the contact lenses cover the entire iris area and extend beyond it in many cases. Given the lens attack images in public databases, one can notice the existence of a lens (see Figure 5.7). This high frequency image information on the edge of the lenses can offer valuable information on the existence and type (textured (attack) or transparent (bona fide)) of a lens. Moreover, in the case of printout attacks, given the high contrast (high frequency) of the iris/sclera boundary, any blurring (frequency reduction) effect of printout attacks can be noticed easier in this area, see examples in Figure 5.8. First, the coarse iris segmentation is performed by using the Iris-Toolkit (USIT) v2.4.2 tool [191] to calculate the pupil and iris boundaries. For the databases captured in an uncontrolled environment, the USIT tool did not perform as expected. For these databases, we deployed the recently proposed Eye Multi-Scale Segmentation (Eye-MMS) [20] lightweight segmentation, as



Attack	Live	Labels
0.974	0.026	Attack
0.588	0.412	Attack
0.897	0.103	Attack
0.150	**0.850**	Live
⋮		
0.989	0.011	Attack
0.993	0.007	Attack
0.993	0.007	Attack

Figure 5.6.: An example illustrating the process of our MSA approach in detail. The iris image with a textured contact lens is from the LivDet-2017-WVU database [243].

will be described in Section 5.4.2. In order to focus on the iris/sclera boundary region, we use the following equation to extend the pupil and iris boundaries:

$$\begin{aligned} r_{inside} &= r_{iris} - ((r_{iris} - r_{pupil}) * s_1) \\ r_{outside} &= r_{iris} + ((r_{iris} - r_{pupil}) * s_2) \end{aligned} \tag{5.4}$$

where r_{iris} and r_{pupil} represent the radiuses of the iris and pupil. s_1 and s_2 determine the extend ratio. In experiment, s_1 and s_2 are both $\frac{2}{5}$. This aims at having an adaptive segmentation to irises of different sizes.

Then, the extended segmentation is normalized using Daugman's rubber sheet expansion technique [62]. The normalized image is of the dimension 512×64 pixels in the experiment. Finally, the Contrast Limited Adaptive Histogram Equalization (CLAHE) algorithm is applied in the normalized image to improve the contrast and enhance the texture. The segmentation and normalization processes are demonstrated in Figure 5.6.

Overlapping micro-Stripes Once the normalized image is obtained, we extract overlapping micro-stripes for two reasons. Our MSA solution is based on an assumption that PA artifacts around the iris/sclera boundary should provide the most useful information to make a PAD decision. Thus, we utilize smaller regions of texture (micro-stripes) to force the classifier to concentrate on iris/sclera boundary area and capture the PA artifacts. On the other hand, the number of training samples available in standard iris databases is limited. In our initial experiment on the full segmentation area (without stripes), we observed that the validation accuracy and loss fluctuate wildly, which indicates that the classifier suffers overfitting as demonstrated by the inferior results and generalization later. Therefore, overlapping micro-stripes leads to lower-dimensional input data and a higher number of training samples, and thus enhance model fitting. Some of these stripes might not contain the information that we are looking for as the iris/sclera boundary and contact lens boundaries occur at different locations (in relation to the segmentation). Moreover, the iris segmentation might not always be precise. Based on these two factors, a fixed stripe is not always optimal for PAD. Therefore, we utilize multiple overlapping stripes. Each of these stripes will produce a PAD decision and the final decision will be the majority vote of these decisions. The superiority of this process will be demonstrated and discussed later in the results section (Section. 5.4.3).

Classification and Fusion Recently, deep learning-based algorithms have been successfully used for various tasks such as segmentation, classification, or object detection. Recent works in the iris PAD field have already attempted to use neural network architectures such as VGG-16 [238], AlexNet [241], or custom CNN [36, 189] based approaches and obtained good performances. However, despite the good detection performance, such solutions demand high computational power due to their network size, which is not acceptable for some operation scenarios. We investigate the possibility of deploying a neural network model with constrained hardware requirements, while retaining high accuracy. In contrast to VGG-16 with 138 million and AlexNet with 60 million parameters, **MobileNetV3-small** [114] only has 2.5 million parameters. On that account, MobileNet V3-small is chosen as the basic structure for our proposed MSA approach to classifying an iris image as a bona fide or attack. We train our own MobileNetV3 from scratch with input stripe size $(512 \times 32 \times 1)$ and use early stopping to reduce overfitting. The training-from-scratch is enabled by our overlapping micro stripes preprocessing, which provides virtually larger training data and lower dimensionality. In the training process, the overlapping stride of 4 pixels and the early stopping patience is 10 in the training process, the maximum training epochs is 100, and the output of our MobileNetV3 network is a score indicating a bona fide or attack decision. In addition, we use Root Mean Square Propagation algorithm (RMSprop) with a learning rate of 0.001 as an optimizer.

Another strategy we exploit for decision-making is majority voting in order to enhance the robustness of our approach. The extraction of overlapping micro-stripes is not only applied for training classifiers but also used for making a final decision. Multiple micro-stripes from each texture image are naturally expected to have the same prediction by the neural network model. Therefore, we sample an odd number of stripes from a normalized iris image with the stride size of 4 pixels instead of using resized iris texture as input in the evaluation stage. In the end, the majority vote is used for the final prediction. The detailed experimental and implementation details are presented later in Section 5.4.2.

Baseline Methods

LBP + SVM: LBP is used for texture-based classification. Such works [97, 174, 255] have used LBP features to train an SVM classifier using a linear kernel. The trained SVM model is used to predict labels (bona fide or attack) in testing databases.
VGG + SVM: VGG [207] is a CNN model for image recognition. VGG16 is now one of the dominant approaches for feature extraction from textual data. Therefore, we use a pre-trained VGG16 to extract features. Then, the Principal Component Analysis (PCA), a linear dimensionality reduction technique, is utilized to project extracted features from each iris image with $7 \times 7 \times 512$ dimension into a 128 dimensional sub-space. Finally, these lower-dimensional features are fed to an SVM with a linear kernel to decide whether the test subject is a bona fide or an attack.
MobileNet V3-small: Considering that our proposed MSA solution uses MobileNet V3-small [114] as a basic network structure, we also train this network from scratch by feeding it the contrast-enhanced full iris images. The network is then used to predict images labels to be either bona fide or attack.

5.4.2. Experimental setup

Database

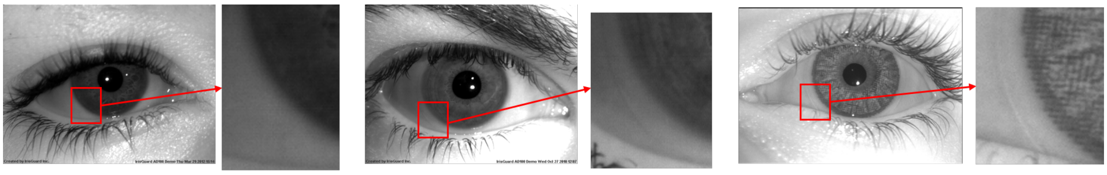
Our proposed approach and the baseline methods are evaluated using three publicly available databases: One is the NDCLD-2015, and the other two databases are from LivDet-Iris 2017 competition, namely the IIITD and WVU databases. The LivDet-Iris 2017 competition [243] included other additional databases (Warsaw, Clarkson and Notre Dame database), however Warsaw and Clarkson are no longer publicly available. Because the Notre Dame database is a subset of NDCLD-2015 database [142] but without the transparent (soft) lens samples, we use the NDCLD-2015 [142] to explore the impact of soft lens on the PAD performance. Table 5.5 presents the attack types and the number of images, capturing devices and environments, respectively. We choose databases for the following reasons: first, the three selected databases were captured by different sensors and under different capturing environment. NDCLD-2015 [142] and IIITD [243] were captured by specific iris sensors under a controlled environment, and WVU was captured by a mobile sensor [241]. Second, WVU was captured in two different environments: indoors (controlled illumination) and outdoors (varying environmental situations), making

Name	Sensor	Capture Env	Type of samples				
			BF	CL	PP	CL+PP	Total
NDCLD-2015	AD100 and LG4000	Controlled	4, 875	2, 425	-	-	7, 300
IIITD	Cognet and Vista	Controlled	2, 250	1, 000	1, 500	1, 500	6, 250
WVU	IriShield MK2120U Mobile	Uncontrolled	702	701	1, 404	1, 402	4, 209

Table 5.5.: Details of the databases. Abbreviations: BF - bona fide, CL + PP - printed images of irises with contact lenses. Other abbreviations are listed in Tab.5.1.

it one of the most challenging databases. Furthermore, these databases contain different types of attacks that can be used for unknown attack scenario validation.

Although NDCLD-2015 [142] provides segmentation information, we perform coarse segmentation on the three databases by the USIT v2.4.2 tool [191] for fairness. Since around 800 images in IIITD and around 850 images in WVU databases failed to have the iris and pupil localized by the USIT tools, we manually segmented these images to be able to sue all iris samples. However, to prove a complete automatic pipeline, we used these manually labeled images to train a deep learning based approach, namely the recently proposed Eye-MMS [20] lightweight segmentation. The Eye-MMS was used to label the complete IIITD and WVU databases. This segmentation was performed on two folds,half (identity-disjoint half) of the manually labeled images to train the Eye-MMS and use it to label the other half, ending up with automatic segmentation of the full database.


(a) No lens (bonafide)

(b) Transparent lens (bonafide)

(c) Textured lens (attack)

Figure 5.7.: Example images of the NDCLD-2015 [142] database. One can notice the subtle difference in the image dynamics around the iris/sclera boundary between different types of lenses and the bona fide (no lens) images. One can notice the visible borders of contact lenses either transparent or textured in the area adjacent to the iris/sclera boundary. Moreover, The transparent lens has smoother border compared to the textured lens that has a rather jagged border.

NDCLD-2015: The 7300 images in the NDCLD-2015 [142] were captured by two sensors, IrisGuard AD100 and IrisAccess LG400. Moreover, all iris images were captured in a windowless indoor lab under NIR illumination. The image resolution of all samples is 640×480 pixels. Example images of this database can be seen in Figure 5.7.

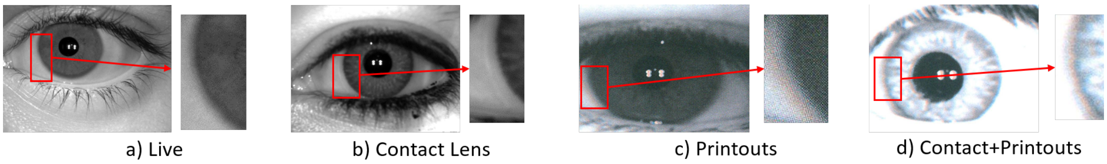


Figure 5.8.: Example images of IIITD [102, 151, 239, 243]. One can notice a border of the zoomed contact lens, and a difference between the bona fide and different attacks. Please note that both printouts attack are noticeably different than the printout attack in the WVU database. Given the high contrast (high frequency) of the iris/sclera boundary, any blurring effect of printout attacks can be noticed more clearly in this area.

IIITD: The images in the LivDet-2017 IIITD were developed using bona fide and textured contact lens iris images from the IIIT-Delhi Contact Lens Iris (CLI) database [151, 239] and print attack images were selected from the IIITD Iris Spoofing (IIS) database [102]. The iris images were captured under controlled environment using two iris sensors. Example images of this database can be seen in Figure 5.8.
WVU: The images in the LivDet-2017 WVU [241, 243] have been acquired in both indoors and outdoors environments. Besides, the outdoor images have been acquired at varying times of the day and weather conditions. The WVU database contains four of the most common presentation: bona fide, textured contact lens, printed iris images, and printed contact lens images. Example images from this database can be seen in Figure 5.9.

IIITD database and WVU database are combined into one database called the IIITD-WVU in the LivDet-Iris 2017 competition, where the IIITD database is used for training and the WVU database is used for testing. Therefore, we will compare this cross-database evaluation with SOTAs and also the winner of the LivDet-Iris 2017 competition in Section 5.4.3.

Implementation details

In this subsection, three different experimental scenarios are identified, and experiments are designed to evaluate the performance between different iris PAD approaches within these scenarios. Because the sensors used in three databases are different from each other, the scenario 2 and scenario 3 below are defined as cross-sensor evaluation.

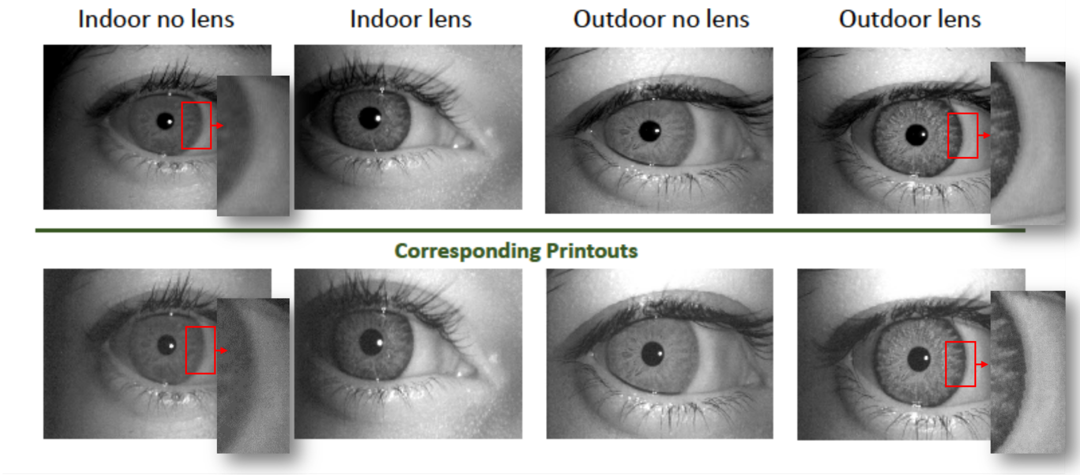


Figure 5.9.: Example images of the WVU [241, 243] database. Similar to the samples of the NDCLD-2015 and the IIITD databases, one can notice the subtle appearance changes in the iris/sclera boundary area in the cases of lens attack and printouts.

Scenario 1: Intra-database experiments

In this scenario, each database is split into 5 cross-validation folds. The subjects in the training and testing partitions are disjoint.

In order to explore the impact of transparent contact lenses on the performance, three experiment compositions of training/testing subsets from NDCLD-2015 are defined:

- The training and testing subset contains both bona fide iris images and textured contact lens attack images.
- The training and testing subset contains bona fide iris images, transparent and textured contact lens attack images. In this way, the images with transparent contact lenses are treated as bona fide iris images.
- No images with transparent lenses are used in the training phase. However, the testing data included both transparent and textured contact lenses, and also bona fide images. Only the samples wearing textured contact should be detected as attack presentation.

These experimental setups followed the one reported in [142]. The experiments on the WVU and IIITD databases are demonstrated separately on different types of attacks and combined attacks.

Scenario 2: Cross-Database experiments

In this experimental setting, we explore the generalizability of iris PAD methods. Cross-database means that one database is used for training a classification model, and other databases are used for testing. Three cases are designed by considering the type of attacks: 1) contact lens attack, 2) printouts and contacts printouts attack, 3) mixed attacks. This scenario is repeated independently five times to replace the 5-folds cross-validation to be consistent with the experimental setup of Scenario 1 and for more accurate and stable performance.

Scenario 3: Cross-Attack (Unknown-Attack) experiments

In a real-world scenario, new types of attacks always appear, which the current iris PAD system has not seen before during training. Thus, the robustness, generalizability, and transferability should also be used to verify the performance of PAD solutions. In this experimentation, we use one type of attack in the training subset, while the testing subsets contain other attacks. For example, a training set has authentic iris images and print attack, while the testing set consists of bona fide iris images, and textured contact lens attacks. Based on our three databases, different compositions are constructed. The reported results use the same setting to scenario 2 (repeat 5 times independently).

Evaluation metrics

The following metrics are used to measure the PAD algorithm performance: CCR, APCER, BPCER and HTER. CCR metric follows the same defined metric in the relative SOTA works [36, 49, 110, 152, 176]. Other works reported the Total Error rate, which is simply $1-CCR$ [238, 241, 242]. In the MSA solution, the threshold for each stripe is 0.5 (which is the network convergence threshold). The binary decisions from multiple stripes are combined in a vote. For baseline methods described in Section 5.4.1, the threshold for the final decision is 0.5 (which is the SVM or network convergence threshold, depending on the method). Moreover, DET curves are plotted to provide a view on the performance on a wider range of operation points (balance between APCER and BPCERvalues). The APCER and BPCER follows the standard definition presented in the ISO/IEC 30107-3 [127].

5.4.3. Results

Scenario 1: Intra-database Results

The results of the intra-database experiments are presented considering different aspects. First, we analyse the contribution of micro-stripes in iris PAD performance based on EER heatmaps. Then, we compare the achieved results with other iris PAD algorithms, including SOTAs and baseline methods described in Section 5.4.1 on each database separately. We also demonstrate the experiments described in Section 5.4.2 to explore the impact of transparent contact lenses on iris PAD algorithms. Finally, the different sizes of micro-stripes (including full texture) are used to examine our MSA solution.

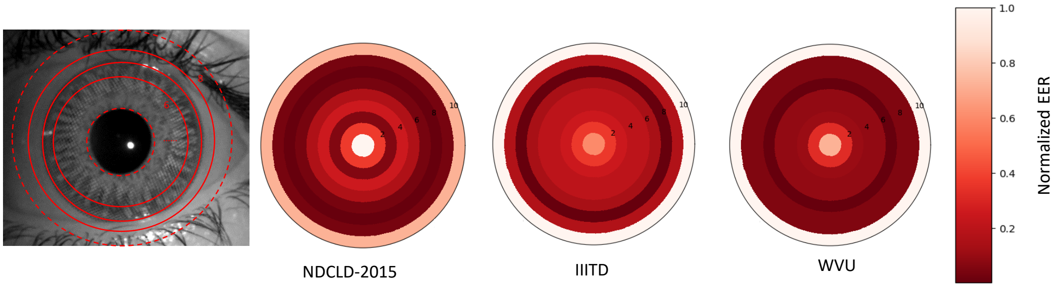


Figure 5.10.: EER heatmaps for different circular rings in the eye region. The area on the outer boundary of the iris and inner boundary of the sclera produces the lowest PAD EER (darker color) over most attacks. This points out the significance of this area for an accurate PAD.

Rationalizing the Micro-stripes: To support our assumption regarding the iris/sclera boundary area significance for PAD decision, we plot the heatmap of each database based on the achieved EER values. This heat map shows the relative significance of different stripes (rings centered around the pupil) by showing the EER achieved if only one stripe is used to make a PAD decision. As shown in Figure 5.10, the region from the inner dash circle to the outer dash circle is divided into nine thinner circular rings. Each circular ring is normalized by Daugman's rubber sheet technique [62] (resulting in our micro-stripe) and used to train a MobileNetV3 network from scratch on the training subset and evaluate on the test subset. Finally, the EER values of all circular rings on each database are normalized between $0-1$ for better relative-visualization to plot as a heatmap. For each database, each stripe (circular ring) was used separately to train an individual network. The trained network was then evaluated on the corresponding stripe data in the testing subset to measure the EER values. The lighter color in the plot refers to the higher EER values, and the darker color corresponds to the lower EER values. In this case, the darker areas are considered to contain more information that aids in PAD decision. In most cases, the region from the fifth to the eighth ring has the darkest red, i.e., the lowest EER values. Besides, the ring closest to the pupil tends to be of light color, which indicates that this part of the iris does not provide much information for a PAD decision. Hence, we believe that the iris/sclera boundary comprises useful information for PAD as its region produces the most accurate PAD decisions.

Metric	Presentation Attack Detection Algorithms (%)								
	LBP [102]	WLBP [255]	DESIST [152]	MH	VGG [207]	MHVF [238]	PSF [49]	MobileNet	MSA(ours)
CCR	74.42	76.98	82.48	85.43	98.92	98.99	98.38	96.22	**99.31**
APCER	6.15	50.58	29.81	21.73	1.54	1.92	-	3.24	0.59
BPCER	38.70	4.41	9.22	9.74	0.78	0.39	-	3.59	0.72

Table 5.6.: Iris presentation attack detection performance (%) of our proposed MSA algorithm and existing algorithms on NDCLD-2015 database.

Comparison with other PAD algorithms: Table 5.6, Table 5.7 and Table 5.8 summarize the results combined with CCR, APCER, and BPCER as mean values across 5-folds cross-validation. The APCER and BPCER are reported at the threshold of 0.5. It should be noticed that the result of NDCLD-2015 here is obtained based on the first experimental setting defined in Section 5.4.2. Because IIITD and WVU database consist of contact lens attack images and printouts attack images, our intra-database experiments demonstrated on the three compositions of databases: 1) detecting contact lens, 2) detecting printouts and contacts printouts, 3) detecting both types of attacks. The three baseline methods, LBP+SVM, VGG+SVM and MobileNet, are described in the Section 5.4.1. Our proposed approach achieved the highest CCR in comparison to SOTA algorithms in the NDCLD-2015 database. Furthermore, in IIITD and WVU database, our MSA solution achieved the best results for contact lens, mixed attack and similar results for the printouts attack detection. Since our MSA method uses the MobileNetV3 as basic network architecture, to demonstrate the superiority of our micro-stripe solution, we use the same network structure and training parameters to classify the full iris image without the micro-stripe processing, which is reported as the "MobileNet" in Table 5.6, Table 5.7 and Table 5.8. We notice that our proposed MSA achieves higher accuracy in comparison to the MobileNet baseline. In comparison to other methods that also used the full image, namely the VGG16 (reported in [238]) and Multi-level Haralick and VGG Fusion (MHVH) [238] in Table 5.6, they achieve higher performance than the MobileNet. This might be due to the large number of parameters and pre-trained nature of the VGG16 solution and the fact that MHVF combines the VGG16 features with handcrafted features. However, Our proposed

MSA significantly outperformed all the previously reported results on the NDCLD15 database. For example, the CCR of the previously best performing solution, the MHVF [238], is increased from 98.99% to 99.31% achieved by our MSA solution. Table 5.8 shows the superiority of our proposed approach in the mobile sensor and uncontrolled capture environment domain by outperforming all the reported baselines on the WVU database, as well as our MobileNet baseline. For example, the best previously reported CCR on the WVU database, by the MobileNet baseline , achieved 95.31%, and our proposed MSA improved that to reach 99.14% to detect contact lens attacks. Considering the properties of our MSA solution, such as image dynamic signs in the iris/sclera boundary area, our methods surpassed all results in all databases for detection of contact lens. In addition, we achieved the best results in mixed attack databases.

Database	Metric	Presentation Attack Detection Algorithms (%)			
		LBP+SVM [102]	VGG+SVM	MobileNet	MSA(ours)
IIITD (CL)	CCR	69.10	97.12	80.12	**98.15**
	APCER	99.10	8.08	54.82	4.88
	BPCER	0.28	0.54	4.23	0.50
IIITD (Print)	CCR	75.96	**99.88**	89.67	99.32
	APCER	23.22	0.06	10.97	0.47
	BPCER	25.11	0.19	9.48	0.95
IIITD (Mixed)	CCR	69.70	97.46	85.80	**98.24**
	APCER	21.60	0.90	7.48	1.91
	BPCER	45.58	5.41	25.99	1.51

Table 5.7.: Iris presentation attack detection performance (%) of our proposed MSA algorithm and existing algorithms with different types of attacks on the database IIITD [116, 157].

Database	Metric	Presentation Attack Detection Algorithms (%)			
		LBP+SVM [102]	VGG+SVM	MobileNet	MSA(ours)
WVU (CL)	CCR	61.50	90.43	95.31	**99.14**
	APCER	36.69	10.81	6.70	0.80
	BPCER	40.05	8.31	2.64	0.94
WVU (Print)	CCR	97.97	98.21	99.10	**99.98**
	APCER	2.53	0.05	1.09	0.00
	BPCER	0.0	8.78	0.13	0.61
WVU (Mixed)	CCR	85.16	92.32	98.22	**99.19**
	APCER	17.00	1.07	0.00	0.35
	BPCER	4.00	40.54	5.56	3.17

Table 5.8.: Iris presentation attack detection performance (%) of our proposed MSA algorithm and existing algorithms with different types of attacks on the database WVU [241, 243].

Impact of Transparent Lens: Baker *et al.* [6] reported that transparent contact lenses lead to the degradation of iris recognition performance. To explore the impact of transparent contact lenses on iris PAD performance, additional experiments are carried out to evaluate if our method can make a correct classification decision unaffected by transparent lenses, i.e., do not classify transparent lenses as attacks and thus reduce the CCR. Three experimental settings are described in Section 5.4.2 and results are shown in Figure 5.11. The average accuracy of three experiments is 99.31%, 99.08%, 99.01%, respectively. The accuracy of employing the model that never sees transparent lenses decreases only 0.07% than the model with already learned related features. This proves that our proposed MSA solution is able to classify the unknown transparent lenses correctly as bona fide.

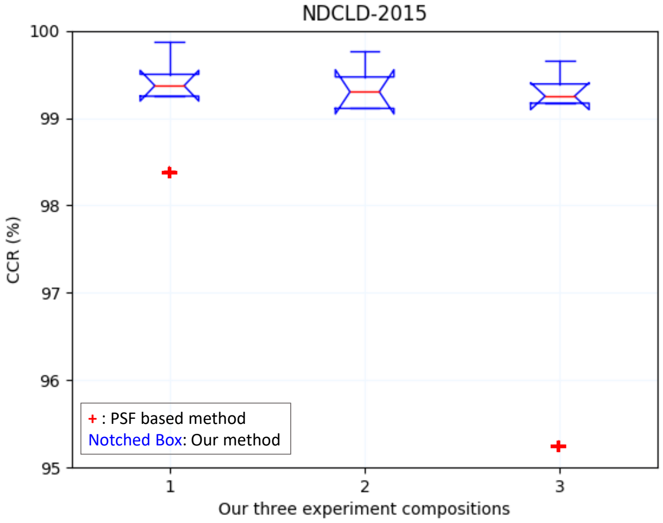


Figure 5.11.: The red cross represents the results from [49], which explore the effect of transparent contact lenses. The notched box describes our cross validation results for three experimental compositions: 1. the training and testing subsets do not have images with transparent lenses, 2. the training and testing subsets contain images with transparent lenses, 3. models trained by subsets without transparent lenses are used to evaluating on testing subsets with transparent lenses.

The red crosses in Figure 5.11 are the average accuracy from the Photometric Stereo Features (PSF) based method reported in [49], where they only reported the first and third experimental setting. Although the same database (NDCLD-2015) is used, they demonstrated in their own constructed two subsets, one called regular, which textured lens has a dot-like pattern, and the other is an irregular subset, without a dot-like pattern. The construction process completely relies on people to distinguish patterns and split them into respective subsets. Since it is hard to replicate the exact experimental settings, we only discuss the performance degradation induced by the transparent lenses on the performance, and we do not compare the CCR values themselves. Their CCR decreased from 98.38% to 95.24%, while ours, from 99.08% to 99.01%, point out higher generalization. Other works that reported on this database did not analyze the issue of transparent lenses and thus are not included in this aspect of the comparison. The WVU and IIITD databases do not include transparent lenses bona fide samples, and thus, we are not able to perform this analyses on these databases.

Impact of Overlapping Micro-Stripes: The purpose of this subsection is to prove the positive effect of our proposed micro-stripes approach. As a baseline, we use the full segmented area as one large stripe with the size of 64×512 pixels, processed by the same network structure and trained with the same experimental settings as the micro-stripes. We also investigate using stripes of different heights (24, 32, and 48 pixels). Figure 5.12 illustrates the CCR achieved by these different settings. Figure 5.12 shows that the MSA approach improved the CCR on all databases, in comparison to the full segmented area classification. Having a micro-stripe of the height of 32 pixels improved the CCR from 94.92% to 99.31% on the NDCLD-2015 database, from 92.15% to 98.14% on the WVU database, and from 94.87% to 99.17% on the WVU database, in comparison to the baseline. We assume that the stripes provide the network with a chance to better generalize on the problem by providing a larger amount of samples, as well as, samples with less complicated information (smaller area). One can imagine the effect of the micro-stripes in a similar manner to data augmentation, leading to lower overfitting. Since we are looking for certain image dynamic signs in the iris/sclera boundary area, and since the segmentation is not always optimal, these signs of attack might occur at different places. The nature of the overlapping micro-stripes is assumed to provide robustness to this localization issue. However, in some scenarios, having very thin micro-stripes might reduce the performance, e.g., 24 pixel stripes on the NDCLD-2015 database (see Figure 5.12). This might be due to the fact that a very small stripe might not have sufficient information to make the detection decision. Therefore, the size 32×512 pixels of the micro-stripe should be considered suitable for higher performance on both databases and both imaging spectrums. This size will be the one reported for all other experiments.

Scenario 2: Cross-Database Results

In this subsection, we demonstrate our proposed method in cross-database where one database is used for training and the other databases are used for testing. In order to

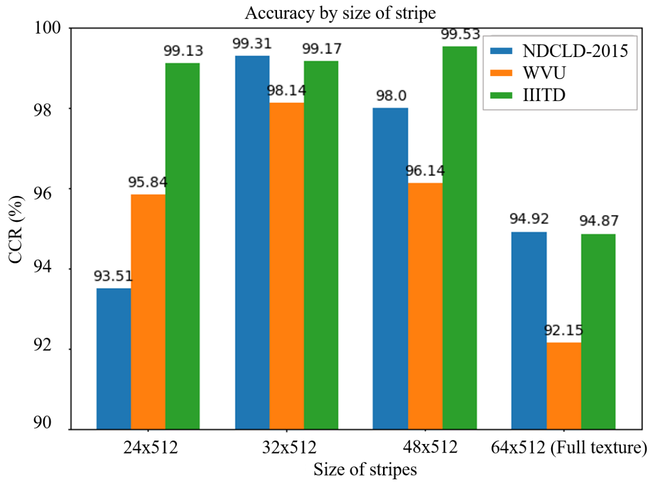


Figure 5.12.: Performance of the different size of overlapping micro-stripes on the three investigated databases. Size of 64×512 means that full normalized iris images are fed to network.

explore the impact of different types of attacks, the following experiments are performed based on contact lens attack (Table 5.9 and Figure 5.14), printouts attack (Table 5.10 and Figure 5.15), and mixed attacks (Table 5.11 and Figure 5.16). Moreover, we also show the EER heatmaps to analyze the contributions of different regions to the iris PAD decisions (See Figure 5.13), and thus further rationalize (see Section 5.4.3) our MSA approach for the generalizability over unknown databases.

First, we look at the heatmaps based on EER values, as shown in Figure 5.13. These heatmaps are generated by using the same approach described in the intra-database scenario (Section 5.4.3). We expand the normal iris segmentation and split it into nine thinner circular rings. The difference is that each ring is trained on one database and tested on another database to perform the cross-database validation and measure the generalizability of the decision produced by each of these rings. By observing the heatmaps, we can notice that the stripes (ring), which are closet to the pupil or too far away from the iris, contribute the least in most cases. In contrast, the stripes from the fifth to the eighth place achieve relatively lower EER values. This is valid for contact lens and printout attacks. These results validate the significance of the iris/sclera boundary region again, this time for cross-database generalization. In the following, we further analyze additional experiments based on CCR values and DET curves separately.

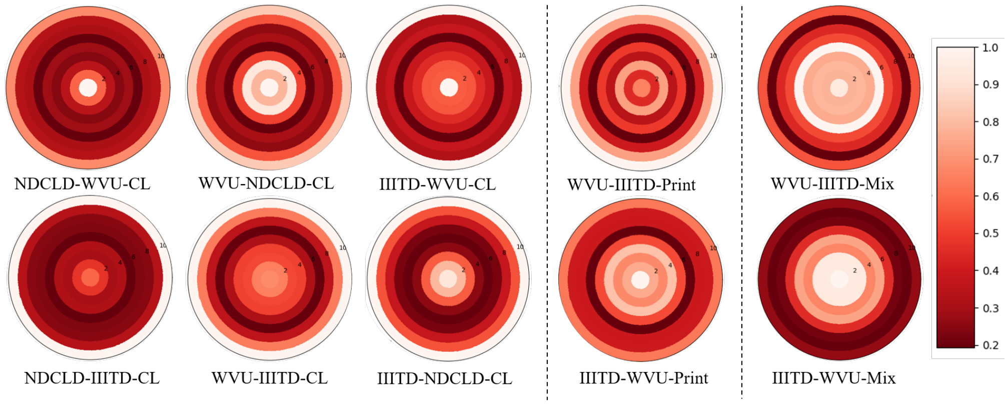


Figure 5.13.: EER-heatmaps of cross-database iris PAD. These EER heatmaps correspond to the three cross-database scenarios, which are divided by dash lines. The first three columns of heatmaps target textured contact lens attack. The fourth column plots are EER values based on the print attack, while the heatmaps in the last column related to the mixed types of attacks (print + textured lenses). The darker the color, the more significant the contribution of this area to the iris PAD result. The labels of the heatmaps indicate the training data, then the testing data, and the testing data attack type (Training Data (NDCLD/WVU/IIITD) - Testing Data (NDCLD/WVU/IIITD) - Testing Attack Type (contact lenses (CL)/printouts (print)/mixed (Mix)).

Contact Lens Attack Detection To find the dynamic signs around iris/sclera boundary area, we use our MSA solution to perform contact lens detection. Table 5.9 shows the evaluation results based on CCR (%) and Figure 5.14 represents the comparison of APCER and BPCER error rates through DET curves. The important observations made from this experiment are as follows: 1) Our MSA solution displays the highest correct classification rate when training on NDCLD-2015 and IIITD databases and testing on other databases. This proves again that using multiple micro stripes improves the performance of contact lens detection even in the cross-database evaluation; 2) VGG+SVM method shows better results than our proposed approach when training in the WVU database. This can be caused by an imbalanced number of training and testing samples. Since the WVU only contains 702 contact lens samples, our MSA model might be overfitted.3) The DET curves of cross contact lens detection indicate that our method has better generalizability than the other three baseline methods, as shown in Figure 5.14.

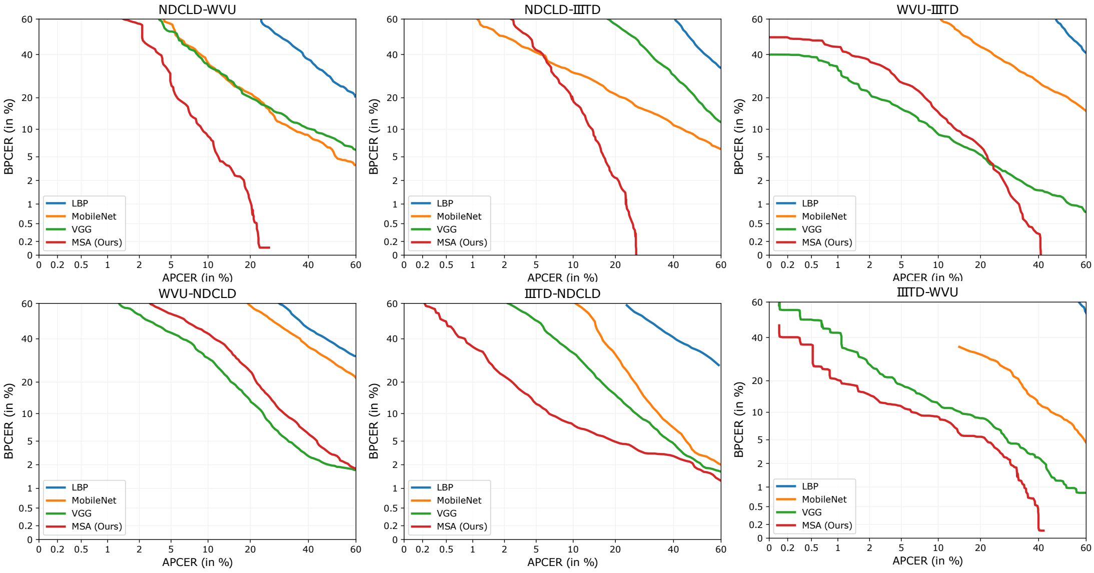


Figure 5.14.: DET curves: Cross-database Performance of Contact Lens

Database		Iris PAD Methods (CCR in %)			
Training	Testing	LBP+SVM	VGG+SVM	MobileNet	Our MSA
NDCLD15	WVU	57.56	79.96	77.90	**91.77**
NDCLD15	IIITD	56.48	75.06	87.13	**91.00**
WVU	NDCLD15	52.39	**84.25**	54.72	58.45
WVU	IIITD	33.60	**91.72**	44.19	69.47
IIITD	WVU	49.93	89.14	78.54	**90.38**
IIITD	NDCLD15	50.47	82.96	79.57	**89.14**

Table 5.9.: Cross-database Performance of Contact Lens.

Printouts Attack Detection Because the NDCLD-2015 database does not contain printouts attack, the cross printout attack is performed when training on the WVU and testing on the IIITD databases and vice versa. It should be noticed that there are two kinds of printouts attacks: 1) bona fide printouts, 2) contact lens printouts. The main observations from this experiment are summarized below:

1. The VGG+SVM method achieves the highest accuracy when using WVU as a training database, while MobileNet baseline obtained the best result in the case of IIITD as a training database, as shown in Table 5.10. Our proposed approach works worse

than VGG+SVM or MobileNet. It seems that micro stripe lost some information in printouts attack samples after expanding the segmentation.

2. The accuracy in both databases is lower than 85%. Compared to cross-database with contact lens with most accuracies higher than 90%, the printout attacks cross-database proved to be a harder problem in our experiments. This is due to the significant difference between printouts attack samples in WVU and IIITD databases. As shown in Figure 5.9 and Figure 5.8, samples in the WVU database have a higher resolution compared to samples in the IIITD database because of various printer types.

3. The DET curves of the four PAD methods for cross printouts attack detection are presented in Figure 5.15.

Database		Iris PAD Algorithms (CCR in %)			
Training	Testing	LBP+SVM	VGG+SVM	MobileNet	Our MSA
WVU	IIITD	42.49	**79.83**	50.10	54.59
IIITD	WVU	59.87	45.86	**83.67**	79.25

Table 5.10.: Cross-database Performance of Printouts and Contacts Printouts.

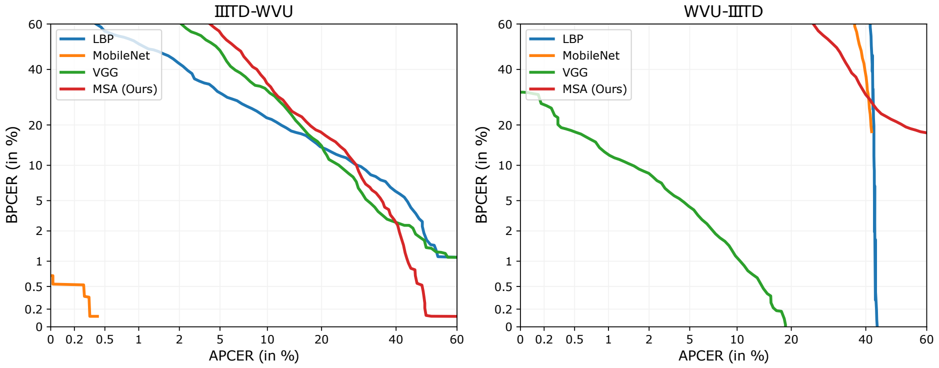


Figure 5.15.: DET curves: Cross-database Performance of Printouts and Contacts Printouts.

Mixed types of Attacks Detection The databases including two types of presentation attack are used to perform the cross-database evaluation. Therefore, the NDCLD-2015 database is removed in this experiment. The IIITD-WVU cross-database evaluation is also used in the LivDet-2017 competition. Thus, our method also compares with the winner and another SOTA algorithm. The main observations from this experiment are summarized below: 1) Our MSA solution performs the best in the IIITD-WVU cross-database evaluation with 93.07% CCR in Table 5.11. In addition, our proposed method achieved lower HTER than the winner of the LivDet-Iris 2017 competition and meta-fusion method [155] as shown in Table 5.12. The results indicate that our MSA approach has better generalizability than SOTA algorithms. 2) VGG+SVM baseline displays a higher CCR in WVU-IIITD composition than the MSA approach. Considering the similar results in the cross-database of printouts attack detection, the problem may still be that the training samples are insufficient for our neural network, resulting in model overfitting. 3) The DET curves of four PAD methods for cross-database are presented in Figure 5.16 and support our stated observations on a wider range of operation points. 4) The heatmaps presented in Figure 5.13 indicates the relatively higher significance of the areas around the iris/sclera boundary in detecting attacks in a cross-database evaluation scenario.

Database		Iris PAD Algorithms (CCR in %)			
Training	Testing	LBP+SVM	VGG+SVM	MobileNet	Our MSA
WVU	IIITD	47.75	**85.74**	72.08	67.77
IIITD	WVU	70.68	70.59	76.45	**93.07**

Table 5.11.: Cross-database Performance of mixed Attacks (Contacts, Printouts, Contacts printouts).

Database		Iris PAD Algorithms (HTER in %)		
Training	Testing	LivDet-Iris 2017 Winner	Meta-Fusion [155]	Our MSA
IIITD	WVU	16.70	14.92	**11.67**

Table 5.12.: Comparison of HTER (%) with LivDet-Iris 2017 and Meta-Fusion approach [155]

Scenario 3: Unknown-Attack Results

In the real-world scenario, a PAD system should be able to handle unknown types of presentation attacks. Therefore, unknown attack detection evaluation is designed using one type of attacks (e.g. printouts) in the training phase and testing with other unseen attacks like contact lenses. There are eight databases combinations for unknown attack detection evaluation as shown in Table 5.13, the corresponding DET curves are represented in Figure 5.17, and the corresponding EER heatmaps are showed in Figure 5.18. The main observations from this experiment are summarized below: 1) Our proposed MSA method

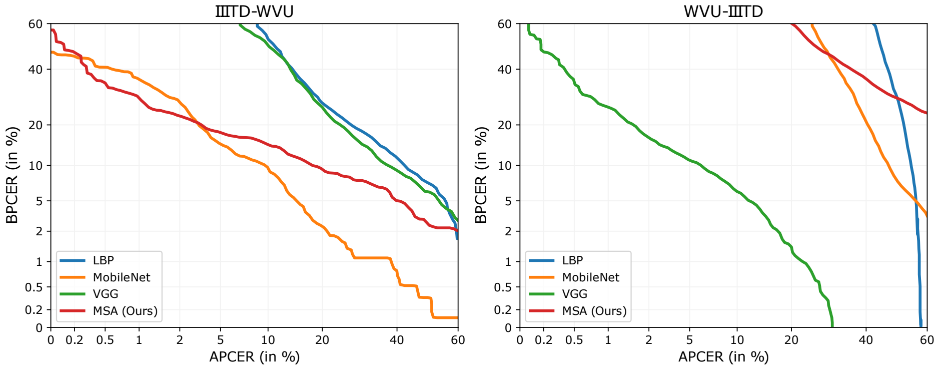


Figure 5.16.: DET curves: Cross-database Performance of Mixed Attacks.

achieves four of the highest accuracy in eight combinations of databases, two combinations are trained with contact lens and tested on printouts, the other two combinations are vice versa. It proved that our method has a better generalization ability than other baselines. 2) In WVU (Print) - NDCLD15 (CL) and WVU (Print) - IIITD (CL) compositions, all methods achieve close to random accuracy. In other compositions, performances are lower than evaluation results in a single database. It indicates that unknown attack detection is a changeling problem and should become a potential criterion to examine the generalizability of PAD systems. 3) The DET curves of four PAD methods for unknown attack detection are presented in Figure 5.17 and support our stated observations on a wider range of operation points. 4) The heatmaps presented in Figure 5.13 indicate the relatively higher significance of the areas around the iris/sclera boundary in detecting attacks in a cross-attack evaluation scenario.

Database		Iris PAD Algorithms (CCR in %)			
Training	Testing	LBP+SVM	VGG+SVM	MobileNet	Our MSA
NDCLD15(CL)	WVU(Print)	59.58	43.98	53.68	**78.16**
NDCLD15(CL)	IIITD(Print)	**86.05**	46.93	47.00	77.14
WVU (CL)	IIITD (Print)	40.02	70.98	71.09	**80.03**
IIITD (CL)	WVU (Print)	14.89	60.85	**88.44**	86.99
WVU(Print)	NDCLD15(CL)	35.52	**56.64**	50.91	40.01
WVU(Print)	IIITD(CL)	54.23	**58.70**	50.57	50.20
IIITD(Print)	WVU(CL)	50.50	50.07	50.07	**64.82**
IIITD(Print)	NDCLD15(CL)	45.83	51.36	53.92	**75.59**

Table 5.13.: Cross-database Performance of Unknown Attacks.

Figure 5.18 shows the EER heatmaps to analyze the contributions of different regions to the iris PAD decisions for unknown attacks, and thus further rationalize (see Section 5.4.3) our MSA approach for the generalizability over unknown attacks. The stripe (ring) close to the pupil contributes very little compared to others, even the outermost stripes. The darker stripes (lower EER values) appear in the fifth to the eighth or ninth place. We conclude that the iris/sclera boundary area has significant contribution to iris PAD decision in the case of cross-attack detection. This is even more obvious when training on textured lenses and testing on print attacks. One possible reason is that the iris samples with the lenses have a clear border as shown in Figures 5.7, 5.8 and 5.9. When we train the model on the print attacks and test on the contact lenses, we find that the contributing stripes are closer to the pupil compared to training on the textures lenses. In the LivDet-Iris 2017 competition, the results suggested that printed iris images are easier to detect than textured contact lenses. Based on this notion, we assume that the print attacks contains more detectable artifacts (blur) distributed over the whole image. In general, the heatmaps indicate that the image dynamics around the iris/sclera boundary contribute relatively more significant information for the PAD of unknown attacks.

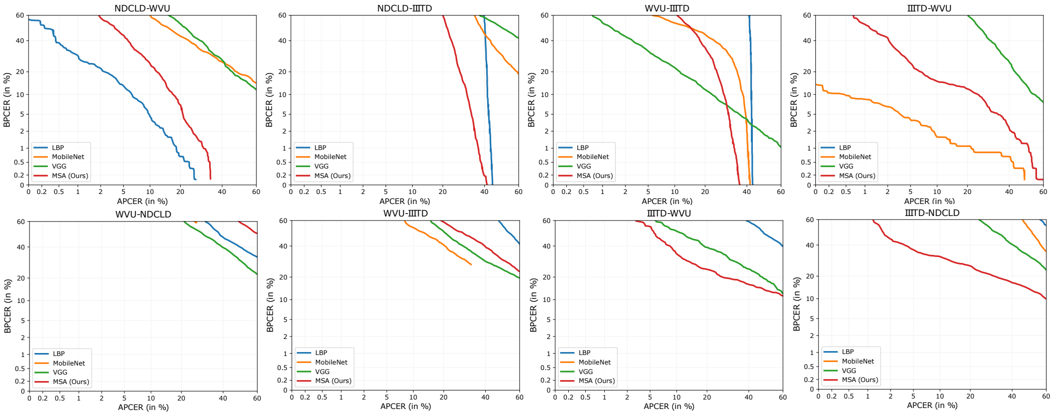
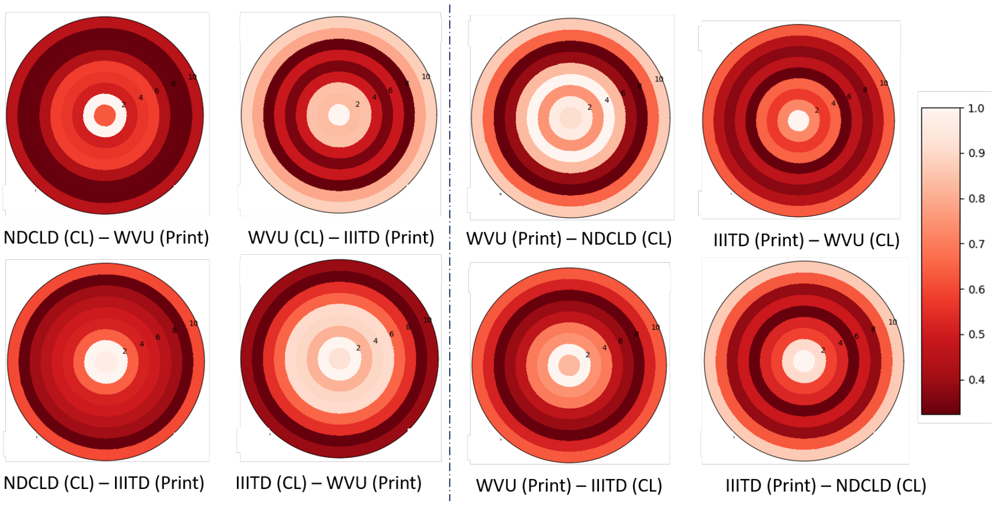


Figure 5.17.: DET curves: Cross-database Performance of Unknown Attacks. The figures in the first row present the situation where the training procedure use contact lenses iris samples and test on the printouts attack. The figures in the second row indicate that using printouts attack samples to train a model and test on the contact lenses attack.



NDCLD (CL) – WVU (Print) WVU (CL) – IIITD (Print) WVU (Print) – NDCLD (CL) IIITD (Print) – WVU (CL)

NDCLD (CL) – IIITD (Print) IIITD (CL) – WVU (Print) WVU (Print) – IIITD (CL) IIITD (Print) – NDCLD (CL)

Figure 5.18.: EER-heatmaps of cross-attack PAD. The plots to the left of the dashed line are the results trained on textured lens attacks and tested on print attacks. In contrast, the plots to the right of the dashed line are the results trained on print attacks and tested on textured lenses. The darker the color, the greater the contribution of this area to the iris PAD result. It can be noticed that the fifth to eighth rings contribute a lot to the iris PAD decisions in most cases. The labels of the heatmaps indicate the training data and type of attack used in training, then the testing data and type of attack used for testing (Training Data (NDCLD/WVU/IIITD) and training attack type (contact lenses (CL)/printouts (print)) - Testing Data (NDCLD/WVU/IIITD) - and testing attack type (CL/print)).

5.4.4. Discussion

The MSA solution [78, 82] was presented to detect iris presentation attacks. The proposed methods focused on the differences (between attack and bona fide) in the image dynamics around the iris/sclera boundary area. To achieve that, multiple thin micro-stripes of the normalized segmentation were extracted to provide more samples and less sample dimensionality, resulting in a better fitted model. The decision of multiple overlapping stripes was fused by a simple majority vote to build the final detection decision. Moreover, the efficiency, robustness, and generalizability of the proposed methods were analyzed, and its results were compared with SOTA iris PAD algorithms in three different scenarios based on two types of attacks: 1) intra-database evaluation, 2) cross-database evaluation,

3) unknown attack detection evaluation. Moreover, the illustrated EER heatmaps based on different pupil-centred rings were used to support the assumption that the iris/sclera boundary contains relatively more significant information for a PAD decision, even in the cases of unknown databases and unknown attacks. The lowest EER values commonly occurred in the fifth to the eighth stripe areas, which were those around the iris/sclera boundary. The experimental results and comparison of the iris PAD algorithms in the intra-database scenario were discussed in detail. The MSA [78, 82] solution surpassed the detection performance of almost all SOTA methods reported on all databases. The CCR of NDCLD-2015 was increased to 99.31%, from 98.99%, in the best performance reported on this database. The same goes for the WVU database, where the CCR was increased from 98.22% to 99.19% in case of mixed printouts and contact lens attacks. Similar to the WVU database, the CCR on the IIITD database was enhanced from 97.46% to 98.24% in mixed types composition. The experiments included a clear rationalization of our micro-stripe process and detailed analyses of the variations in the micro-stripe size. In addition, we demonstrated that our MSA solution does not demonstrate the common issue of confusing bona fide transparent lenses with attack textured lenses. Moreover, the proposed MSA approach can be deployed on low-end devices as the model only has 2.5 million parameters.

In the cross-database evaluation scenario, it can be observed that the proposed method performs better than other methods in case of contact lens attacks when there are sufficient training samples. Moreover, the MSA solution outperformed the winner of the LivDet-2017 competition and another recent SOTA algorithm in the IIITD-WVU composition database. In the scenario of unknown attack detection, it can be concluded that detecting unseen attacks is a challenging task based on the observed large decrease in performance compared to intra-database evaluation. Nevertheless, the proposed MSA method achieved the four highest accuracies in eight combinations of experiments. VGG+SVM method obtained the two best results, while the LBP+SVM method and MobileNet method had only one of the highest accuracy, respectively. Therefore, the presented method had better generalizability. Though no single method outperformed in all experimental setups, the results indicated that deep learning-based methods were more accurate than hand-crafted features-based methods.

5.5. Attention-based iris PAD via pixel-wise supervision

As introduced in 5.2, most results obtained by iris PAD solutions were reported under intra-database scenarios and it is unclear if such solutions can generalize well across databases and capture spectra. One possible reason that these PAD methods ran the risk of overfitting is the binary label supervision during the network training, which serves global information learning but weakens the capture of local discriminative features. This section presents an attention-based deep pixel-wise binary supervision (A-PBS) method [79]. A-PBS utilizes pixel-wise supervision to capture the fine-grained pixel/patch-level cues and attention mechanism to guide the network to automatically find regions where most contribute to an accurate PAD decision. Extensive experiments are performed on six NIR and one visible-light iris databases to show the effectiveness and robustness of proposed A-PBS methods. Extensive experiments are additionally conducted under intra-/cross-database and intra-/cross-spectrum for detailed analysis. The results of the experiments indicates the generalizability of the proposed A-PBS iris PAD approach, as a response to *RQ8*.

5.5.1. Methodology

This section starts by introducing the DenseNet [117], which is used as a preliminary backbone architecture. Then, the Pixel-wise Binary Supervision (PBS) and Attention-based PBS (A-PBS) methods are described. We presented this approach initially in [79], however, we extend it here by investigating its advantages on different attack types, iris images captured in the visible spectrum, and cross-spectrum deployments. Figure 5.19 depicts an overview of our different methods. The first gray block (a) presents the basic DenseNet architecture with binary supervision, the second gray block (b) introduces the binary and PBS, and the third block (c) is the PBS with the fused multi-scale spatial attention mechanism (A-PBS).

Baseline: DenseNet DenseNet [117] presented direct connection between any two layers with the same feature-map size in a feed-forward fashion. The reasons inspiring our selection of DensetNet are: 1) DenseNets naturally integrate the properties of identity mappings and deep supervision following a simple connectivity rule. 2) DenseNet has already demonstrated its superiority in iris PAD [61, 205, 242]. Figure 5.19.(a) illustrates that we reuse two dense and transition blocks of pre-trained DenseNet121. An average pooling layer and a Fully-connected (FC) classification layer are sequentially appended, following the second transition block, to generate the final prediction to determine whether

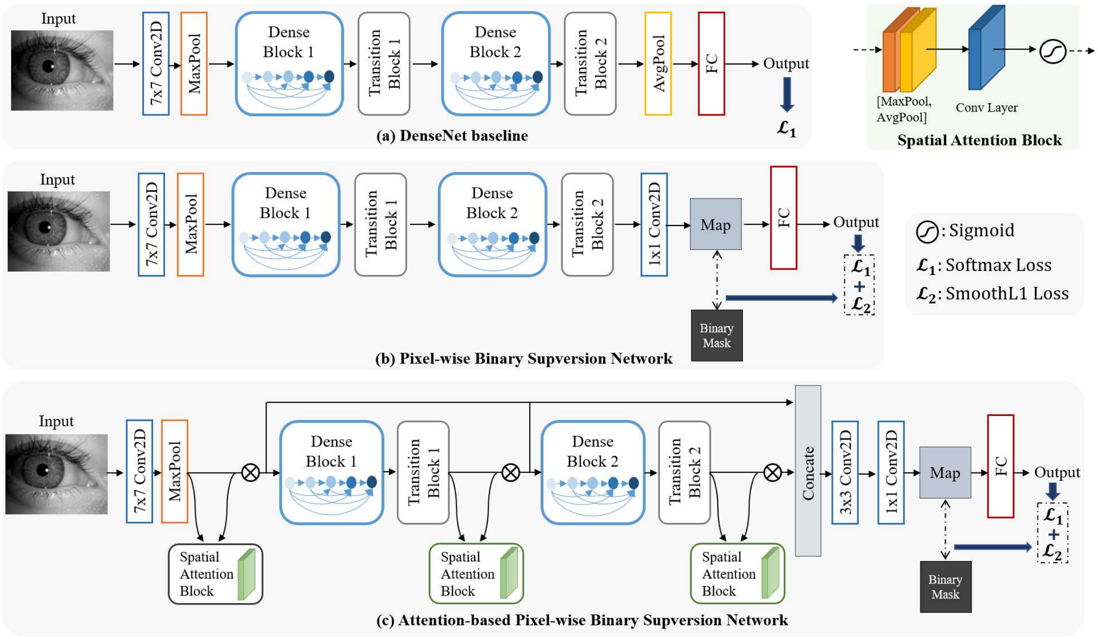


Figure 5.19.: An overview of (a) baseline DenseNet, (b) proposed PBS and (c) proposed A-PBS networks.

the iris image is bona fide or attack. PBS and A-PBS networks are extended on this basic architecture later.

Pixel-wise Binary Supervision Network (PBS) By reviewing the recent iris PAD literature [77, 82, 155, 205], it can be found that CNN-based methods outperformed hand-crafted feature-based methods. In typical CNN-based iris PAD solutions, networks are designed such that feeding pre-processed iris image as input to learn discriminative features between bona fide and artifacts. To that end, a FC layer is generally introduced to output a prediction score supervised by binary label (bona fide or attack). Recent face PAD works have shown that auxiliary supervision [83, 93, 168] achieved significant improvement in detection performance. Binary label supervised classification learns semantic features by capture global information but may lead to overfitting. Moreover, such embedded 'globally' features might lose the local detailed information in spatial position. These drawbacks give us the insight that adding pixel-wise binary along with binary supervision might improve the PAD performance. First, such supervision approach can be seen as a combination of patch-based and vanilla CNN based methods. To be specific, each pixel-wise score in output feature map is considered as the score generated from the patches in an iris image. Second, the binary mask supervision would be provided for the deep embedding features in each spatial position. Figure 5.19.(b) illustrates the network details that an intermediate feature map is produced before the final binary classification layer. The output from the *Transition Block 2* is 384 channels with the map size of 14×14. A 1×1 convolution layer is added to produce the intermediate map. In the end, an FC layer is utilized to generate a prediction score.

Attention-based PBS Network (A-PBS) The architecture of PBS is designed coarsely (simply utilizing the intermediate feature map) based on the DenseNet [117], which might be sub-optimal for iris PAD task. To enhance that, and inspired by Convolutional Block Attention Mechanism (CBAM) [236] and MLF [77], we propose an A-PBS method with multi-scale feature fusion (as shown in Figure 5.19.(c)). Even though PBS boosts iris PAD performance under intra-database/-spectrum, it shows imperfect invariation under more complicated cross-PA, cross-database, and cross-spectrum scenarios (See results in Table 5.17, 5.23, and 5.22). As a result, it is worth finding the important regions to focus on, although it contradicts learning *more* discriminative features. In contrast, the attention mechanism aims to automatically learn *essential* discriminate features from inputs that are relevant to PA detection. Woo *et al.* [236] presented an attention module consisting of the channel and spatial distinctive sub-modules, which possessed consistent improvements in various classification and detection tasks across different network architectures. Nonetheless, only spatial attention module is employed in our case due to the following reasons. The first reason is that the Squeeze-and-Excitation (SE) based channel attention module focuses only on the inter-channel relationship by using dedicated global feature descriptors. Such channel attention module may lead to a loss of information (e.g., class-deterministic pixels) and may result in further performance degradation when the domain is shifted, e.g., different sensors and changing illumination. Second, a benefit of the spatial attention module is that the inter-spatial relationship of features is utilized. Specifically, it focuses on *'where'* is an informative region, which is more proper for producing intermediate feature maps for supervision. Furthermore, based on the fact that the network embeds different layers of information at different levels of abstraction, the MLF [77] approach confirmed that the fusing deep feature from multiple layers is beneficial to enhance the robustness of the networks in the iris PAD task. Nevertheless, we propose to fuse feature maps generated from different levels directly within the network instead of fusing features extracted from a trained model in MLF [77], because finding the best combination of network layers to fuse is a challenging task and difficult to generalize well, especially when targeting different network architectures.

Figure 5.19 illustrates that three spatial attention modules are appended after *MaxPool*, *Transition Block 1*, and *Transition Block 2*, respectively. The feature learned from the *MaxPool* or two *Transition Blocks* can be considered as low-, middle- and high-level features and denoted as

$$\mathcal{F}_{level} \in \mathbb{R}^{C\times H\times W}, \quad level \in \{low, mid, high\}\ . \tag{5.5}$$

Then, the generated attention maps $\mathcal{A}_{level} \in \mathbb{R}^{H\times W}$ encoding where to emphasize or suppress are used to refine $\mathcal{F}_{level}$. The refined feature $\mathcal{F}'_{level}$ can be formulated as $\mathcal{F}'_{level} = \mathcal{F}_{level} \otimes \mathcal{A}_{level}$ where $\otimes$ is matrix multiplication. Finally, such three different level refined features are concatenated together and then fed into a 1×1 convolution layer to produce the pixel-wise feature map for supervision. It should be noticed that the size of convolutional kernel in three spatial attention modules is different. As mentioned earlier, the deeper the network layer, the more complex and abstract the extracted features. Therefore, we should use smaller convolutional kernels for deeper features to locate useful region. The kernel sizes of low-, middle- and high-level layers are thus set to 7, 5, and 3, respectively. The experiments have been demonstrated later in Section 5.5.2 and showed that in most experimental setups, the A-PBS solution exhibited superior performance and generalizability in comparison to the PBS and DenseNet approaches.

Loss Function

In the training phase, BCE loss is used for final binary supervision. For the sake of robust PBS needed in iris PAD, Smooth L1 (SmoothL1) loss is utilized to help the network reduce its sensitivity to outliers in the feature map. The equations for SmoothL1 is shown below:

$$\mathcal{L}_{SmoothL1} = \frac{1}{n}\sum z\ , where \quad z = \begin{cases} \frac{1}{2}\cdot(y-x)^2, & if \quad |y-x| < 1 \\ |y-x| - \frac{1}{2}, & otherwise \end{cases} \tag{5.6}$$

n is the amount number of pixels in the output map (14×14 in our case). The equation of BCE is:

$$\mathcal{L}_{BCE} = -[y \cdot \log p + (1-y)\cdot \log(1-p)]\ , \tag{5.7}$$

where y in both loss equations presents the ground truth label. x in SmoothL1 loss presents to the value in feature map, while p in BCE loss is predicted probability. The overall loss $\mathcal{L}_{overall}$ is formulated as $\mathcal{L}_{overall} = \lambda \cdot \mathcal{L}_{SmoothL1} + (1-\lambda)\cdot \mathcal{L}_{BCE}$. In our experiments, the λ is set to 0.2.

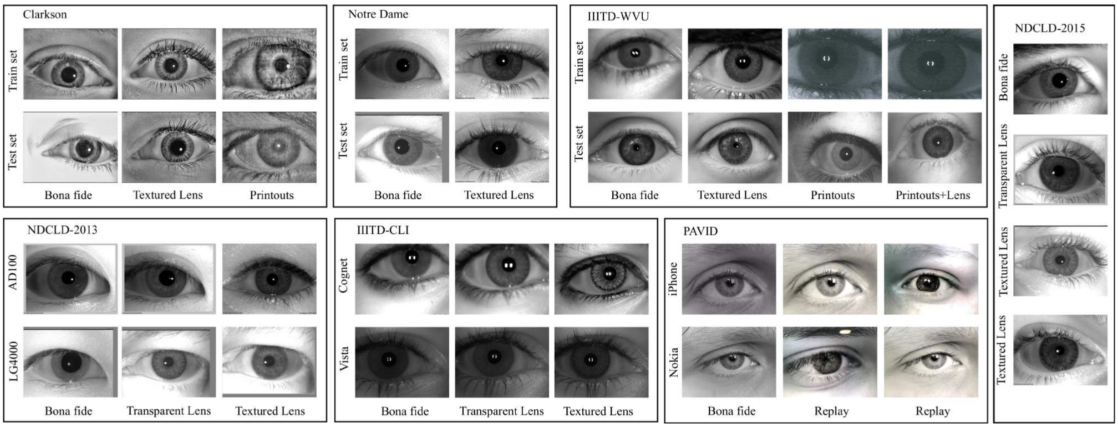


Figure 5.20.: Iris image samples from the used databases. It should be noted that transparent lens is classified as bona fide in our case. Only PAVID database was captured under the visible spectrum.

5.5.2. Experimental setup

Databases

The DenseNet, PBS, and A-PBS were evaluated on multiple databases: three NIR-based databases comprising of textured contact lens attacks captured by different sensors [142, 143, 239], and three databases (Clarkson, Notre Dame and IIITD-WVU) from the LivDet-Iris 2017 competition [243] (also NIR-based). The Warsaw database in the LivDet-Iris 2017 is no longer publicly available due to General Data Protection Regulation (GDPR) issues. For the experiments on NDCLD13, NDCLD15, IIIT-CLI databases, 5-fold cross-validation was performed due to no pre-defined training and testing sets. For the experiments in competition databases, we followed the defined data partition and experimental setting [243]. In addition to above NIR-based iris databases, we also perform experiments on another publicly available database where images were captured under the visible spectrum, named Presentation Attack Video Iris Database PAVID [190]. Subjects in each fold or defined partition are dis-joint. The image samples can be found in Figure 5.20 and the summery and detailed description of the used databases can be found in Chapter 2.

Implementation Details

In the training phase, we performed class balancing by under-sampling the majority class for the databases, whose distribution of bona fides and attacks are imbalanced in the training set. Data augmentation was performed during training using random horizontal flips with a probability of 0.5. The model weight of DenseNet, PBS and A-PBS models were first initialized by the base architecture DenseNet121 trained on the ImageNet dataset and then fine-tuned by iris PAD data, by considering the limited amount of iris data. The Adam optimizer was used for training with a initial learning rate of $1e^{-4}$ and a weight decay of $1e^{-6}$. To further avoid overfitting, the model was trained with the maximum 20 epochs and the learning rate halved every 6 epochs. The batch size is 64. In the testing stage, the binary output was used as a final prediction score. The proposed method was implemented using the Pytorch.

Evaluation Metrics

The following metrics are used to measure the PAD algorithm performance: APCER, BPCER, and HTER as introduced in Chapter 2 and were adopted in most PAD literature including in LivDet-Iris 2017. The threshold for determining the APCER and BPCER is 0.5 as defined in the LivDet-Iris 2017 protocol. In addition, for further comparison with the state-of-the-art iris PAD algorithms on IIITD-CLI [151, 239] database, we also report the CCR. Furthermore, to enable the direct comparison with [205], we evaluate the performance of our presented DenseNet, PBS, and A-PBS methods in terms of TDR at a false detection rate of 0.2%, as [205] claims that this threshold is normally used to demonstrate the PAD performance in practice. TDR is 1 -APCER, and false detection rate is defined to be the same as BPCER, we therefore use BPCER. An EER locating at the intersection of APCER and BPCER is also reported under cross-database and cross-spectrum settings (results as shown in Table 5.17, 5.22, and 5.23). The metrics beyond APCER and BPCER are presented to enable a direct comparison with reported results in state-of-the-arts.

5.5.3. Results

This section presents the experimental results from two aspects: intra-sepcturm (in NIR, and in visible) and cross-database evaluation, and cross-spectrum evaluation.

Intra-spectrum and cross-database evaluation results

This section presents the evaluation results on different databases and comparison to state-of-the-art algorithms. The comparison to state-of-the-arts depends mainly on the reported results in the literature, as most algorithms are not publicly available or their technical description is insufficient to ensure error-free re-implementation. Therefore, we aim to report the widest range of metrics used in other works to enable an extensive comparison. First, the results from different aspects/metrics on LivDet-Iris 2017 database are reported in Table 5.14 which compare our solution with state-of-the-art PAD methods, 5.15 that report the results in terms of TDR at low BPCER, and 5.16 that investigate the performance on different PAs. Then, we demonstrate the experiments under cross-database scenarios by using the three databases in LivDet-Iris 2017 competition to verify the generalizability of our A-PBS solution. Furthermore, the results on NDCLD-2013/NDCLD-2015 and IIITD-CLI databases are presented in Table 5.18, 5.19 and Table 5.20, respectively. We further perform the experiment on the PAVID database in visible spectrum (results in Table 5.21). In this section, we also provide explainability analyses using attention map visualisations for further visual reasoning of the presented solution.

Iris PAD in the NIR spectrum The results on each databases, LivDet-Iris 2017 Databases, NDCLD-2013/2015 Database, and IIITD-CLI Database collected in the NIR spectrum will analyzed first. Then, the results on one database PAVID invisible spectrum will be presented.

Results on the LivDet-Iris 2017 Databases: Table 5.14 summarizes the results in terms of APCER, BPCER, and HTER on the LivDet-Iris 2017 databases. We evaluate the algorithms on databases provided by LivDet-Iris 2017. The evaluation and comparison on LivDet-Iris 2020 are not included due to 1) no officially offered training data, 2) not publicly available test data. Moreover, LivDet-Iris 2017 databases are still considered as a challenging task, because the experimental protocols are designed for complicated cross-PA and cross-database scenarios. In this chapter, we aim to focus on the impact of the algorithm itself on PAD performance rather than the diversity of data. Consequently, to make a fair comparison with state-of-the-art algorithms on equivalent data, we compare to the Scratch version of the D-NetPAD results [205], because Pre-trained and Fine-tuned D-NetPAD used additional data (including part of Notre Dame test data) for training. This was not an issue with the other compared state-of-the-art methods.

It can be observed in Table 5.14 that A-PBS architecture achieves significantly improved performance in comparison to DenseNet and also slightly lower HTER values than the PBS model in all cases. For instance, the HTER value on Notre Dame is decreased from 8.14% by DenseNet and 4.97% by PBS to 3.94% by A-PBS. Although the slightly worse results on Notre Dame might be caused by the insufficient data in the training set, our PBS and A-PBS methods show significant superiority on the most challenging IIITD-WVU database. Moreover, Figure 5.21 illustrates the PAD score distribution of the bona fide and PAs for further analysis. The PAD score distribution generated by A-PBS shows an evident better separation between bona fide (green) and PAs (blue). In addition to reporting the results determined by a threshold of 0.5, we also measure the performance of DenseNet, PBS,

Database	Metric	Winner [243]	SpoofNet [148]	Meta-Fusion [155]	D-NetPAD [205]	MLF [77]	MSA [78, 82]	DenseNet	PBS	A-PBS
Clarkson	APCER	13.39	33.00	18.66	5.78	-	-	10.64	8.97	6.16
	BPCER	0.81	0.00	0.24	0.94	-	-	0.00	0.00	0.81
	HTER	7.10	16.50	9.45	**3.36**	-	-	5.32	4.48	**3.48**
Notre Dame	APCER	7.78	18.05	4.61	10.38	2.71	12.28	16.00	8.89	7.88
	BPCER	0.28	0.94	1.94	3.32	1.89	0.17	0.28	1.06	0.00
	HTER	4.03	9.50	**3.28**	6.81	**2.31**	6.23	8.14	4.97	3.94
IIITD-WVU	APCER	29.40	0.34	12.32	36.41	5.39	2.31	2.88	5.76	8.86
	BPCER	3.99	36.89	17.52	10.12	24.79	19.94	17.95	8.26	4.13
	HTER	16.70	18.62	14.92	23.27	15.09	11.13	10.41	**7.01**	**6.50**

Table 5.14.: Iris PAD performance of our presented DenseNet, PBS, and A-PBS solutions, and existing state-of-the-art algorithms on LivDet-Iris 2017 databases in terms of APCER (%), BPCER (%) and HTER (%) which determined by a threshold of 0.5. The *Winner* in first column refers to the winner of each competition database. Bold numbers indicate the two lowest HTERs.

Database		TDR (%) @ 0.2% BPCER			
		D-NetPAD [205]	DenseNet	PBS	A-PBS
Clarkson		92.05	92.89	**94.02**	92.35
Notre Dame	K	**100.00**	99.68	99.78	99.78
	U	66.55	58.33	76.89	**90.00**
IIITD-WVU		29.30	58.97	69.32	**72.00**

Table 5.15.: Iris PAD performance reported in terms of TDR (%) at 0.2% BPCER on the LivDet-Iris 2017 databases. K indicates known test subset and U is unknown subset. The highest TDR is in bold.

and A-PBS in terms of its TDR at 0.2% BPCER (to follow state-of-the-art trends [205]) in Table 5.15. It is worth noting that our A-PBS method achieves the highest TDR value (90.00%) on unknown-test set in Notre Dame, while the second-highest TDR is 76.89% achieved by PBS.

Furthermore, we explore the PAD performance based on each presentation attack in LivDet-Iris 2017 database [243]. Because the Notre Dame database contains only textured contact lenses, we report the results on Clarkson and IIITD-WVU databases in Table 5.16. The results show that textured contact lens attacks obtain higher APCER values than printouts attack in most cases, e.g., the APCER value on textured lens attack is 10.59% and on printouts is 1.52% both achieved by A-PBS solution. Hence, we conclude that contact lens is more challenging than printouts in most cases.

In addition to intra-dataset evaluation, we further evaluate the generalizability of our models under cross-database scenario, e.g., the model trained on Notre Dame is tested on Clarkson and IIITD-WVU. As shown in Table 5.17, the A-PBS model outperforms DenseNet and PBS in most cases, which verifying that additional spatial attention modules can

Database	Clarkson			IIITD-WVU			
# Images	# 1485	# 908	# 765	# 704	# 1404	# 701	# 1402
Metric	BPCER	APCER (PR)	APCER (CL)	BPCER	APCER (PR)	APCER (CL)	APCER (PR-CL)
DenseNet	0.00	0.66	22.48	17.95	3.06	8.27	0.00
PBS	0.00	0.44	19.08	8.26	11.68	5.42	0.00
A-PBS	0.81	1.32	10.59	4.13	11.68	17.97	0.86

Table 5.16.: Iris PAD performance reported based on each presentation attack on the LivDet-Iris-2017 database in terms of BPCER (%) and APCER (%). PR, CL and PR-CL refer to printouts,textured contact lens, printed textured contact lens, respectively. The Notre Dame database is omitted because it comprises only texture contact lens attack and the results are the same as in Table 5.14. It can be observed that textured contact lens attack is more challenging than printouts in most cases.

reduce the overfitting of the PBS model and capture fine-grained features. Furthermore, the DenseNet and A-PBS models trained on Notre Dame even exceed the prior state-of-the-arts when testing on the IIIT-WVU database (8.81% HTER by DenseNet and 8.95% by A-PBS, while the best prior state-of-the-art achieved 11.13% (see Table 5.14)). It should be noted that the APCER values on Notre Dame are significant higher by using models either trained on Clarkson or IIITD-WVU. Because Notre Dame training dataset contains only textured lens attacks while Clarkson and IIIT-WVU testing datasets comprise of both textured lens and printouts attacks, which makes this evaluation scenario partially consider unknown PAs. In such an unknown-PAs situation, our A-PBS method achieved significantly improved results. In general, the cross-database scenario is still a challenging problem since many D-EER values are above 20% (Table 5.17).

Results on the NDCLD-2013/2015 Databases: Table 5.18 compares the iris PAD performance of our models with five state-of-the-art methods on NDCLD-2015 and two different subsets in the NDCLD-2013 database. It can be seen from Table 5.18 that our A-PBS model outperformed all methods on all databases, revealing the excellent effectiveness of a combination of PBS and attention module on textured contact lens attacks. In addition to comparison with state-of-the-art algorithms, we also report the TDR (%) at 0.2% BPCER in Table 5.19. It can be found that despite all three models produce similarly good results, A-PBS obtains slightly better performance than DenseNet and PBS. The near-perfect results on NDCLD-2013/-2015 databases hint at the obsolescence and limitations of the current iris PAD databases and call for the need for more diversity in iris PAD data.

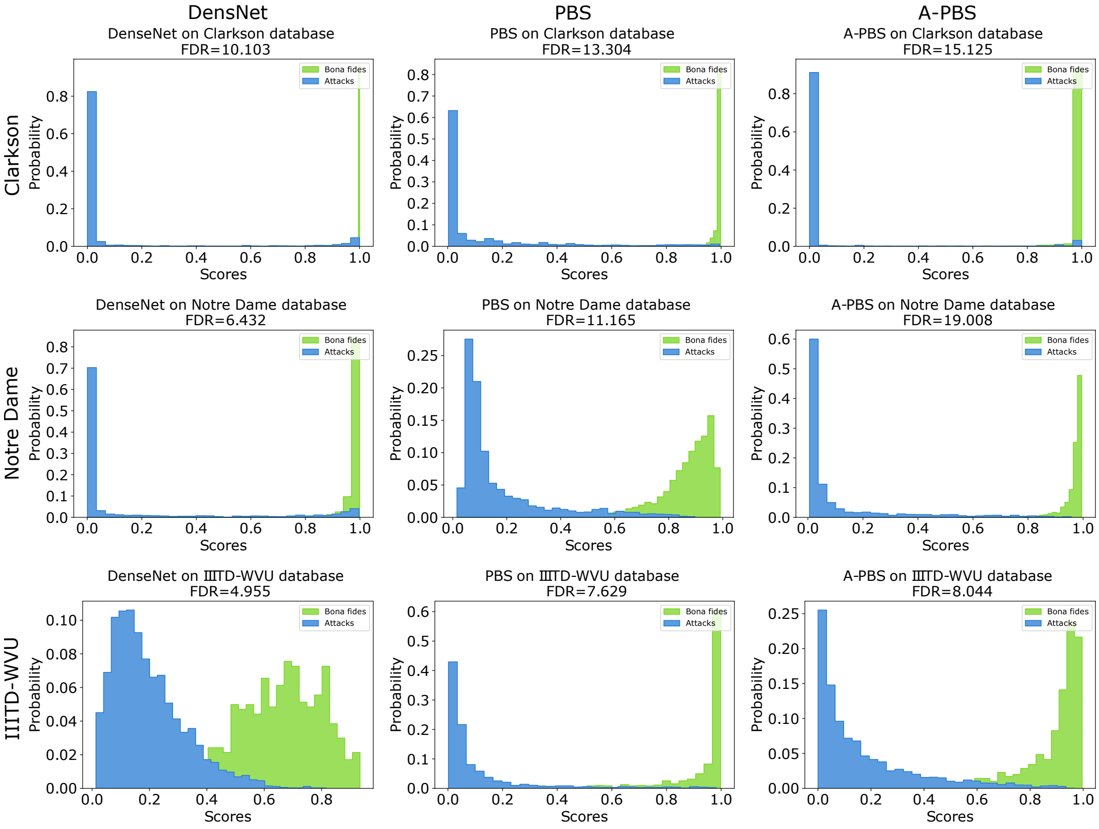


Figure 5.21.: PAD score distribution of bona fide (green) and PAs (blue) on the LivDet-Iris 2017 databases. The histogram top to bottom are results on Clarkson, Notre Dame and IIITD-WVU databases, and the histograms from left to right are produced by DenseNet, PBS, and A-PBS, respectively. The larger separability (measured by Fisher discriminant ratio [57, 171]) and smaller overlap indicate higher classification performance. It can be observed that the our proposed A-PBS method achieved the highest FDR value on all three databases.

Train dataset	Notre Dame							
Test dataset	Clarkson				IIITD-WVU			
Metric	EER	HTER	ACPER	BPCER	EER	HTER	ACPER	BPCER
DenseNet	30.43	32.01	51.29	**12.73**	7.84	**8.81**	**5.93**	11.69
PBS	48.36	47.28	28.15	66.4	15.52	14.54	22.24	6.83
APBS	**20.55**	**23.24**	**14.76**	31.72	**6.99**	8.95	15.34	**2.56**
Train dataset	Clarkson							
Test dataset	Notre Dame				IIITD-WVU			
Metric	EER	HTER	ACPER	BPCER	EER	HTER	ACPER	BPCER
DenseNet	22.33	31.11	62.22	0.00	26.78	42.40	84.80	0.00
PBS	28.61	32.42	64.83	0.00	25.78	42.48	84.97	0.00
APBS	**21.33**	**23.08**	**46.16**	0.00	**24.47**	**34.17**	**68.34**	0.00
Train dataset	IIITD-WVU							
Test dataset	Notre Dame				Clarkson			
Metric	EER	HTER	ACPER	BPCER	EER	HTER	ACPER	BPCER
DenseNet	18.28	19.78	36.56	3.00	22.64	48.55	**0.00**	97.10
PBS	**12.39**	**16.86**	**33.33**	0.39	37.24	47.17	**0.00**	94.34
APBS	15.11	27.61	54.72	**0.33**	**21.58**	**21.95**	20.80	**32.10**

Table 5.17.: Iris PAD performance measured under cross-database scenarios and reported in terms of EER (%), HTER (%), APCER (%), and BPCER (%). APCER and BPCER are determined by a threshold of 0.5. The lowest error rate is in bold.

Database	Metric	Presentation Attack Detection Algorithm (%)							
		LBP[102]	WLBP [255]	DESIST [152]	MHVF [238]	MSA [78, 82]	DenseNet	PBS	A-PBS
NDCLD-2015 [142]	ACPER	6.15	50.58	29.81	1.92	0.18	1.58	1.09	0.08
	BPCER	38.70	4.41	9.22	0.39	0.00	0.14	0.00	0.06
	HTER	22.43	27.50	19.52	1.16	0.09	0.86	0.54	**0.07**
NDCLD13 (LG4000) [143]	APCER	0.00	2.00	0.50	0.00	0.00	0.20	0.00	0.00
	BPCER	0.38	1.00	0.50	0.00	0.00	0.28	0.03	0.00
	HTER	0.19	1.50	0.50	**0.00**	**0.00**	0.24	0.02	**0.00**
NDCLD13 (AD100) [143]	APCER	0.00	9.00	2.00	1.00	1.00	0.00	0.00	0.00
	BPCER	11.50	14.00	1.50	0.00	0.00	0.00	0.00	0.00
	HTER	5.75	11.50	1.75	0.50	0.50	**0.00**	**0.00**	**0.00**

Table 5.18.: Iris PAD performance of our proposed methods and existing state-of-the-arts on NDCLD-2013/-2015 databases with a threshold of 0.5. The best performance (in terms of lowest HTER) is in bold

Results on the IIITD-CLI Database: Since most of the existing works reported the results using CCR metric on IIITD-CLI database [151, 239], we also strictly follow its experimental protocol where we show the experimental results in Table 5.20. In addition to CCR, the

Database	TDR (%) @ 0.2% BPCER		
	DenseNet	PBS	A-PBS
NDCLD-2015	99.45	99.84	**99.96**
NDCLD13 (LG4000)	99.75	**100.00**	**100.00**
NDCLD13 (AD100)	100.00	100.00	100.00
IIITD-CLI (Cognet)	99.02	**99.59**	99.57
IIITD-CLI (Vista)	100.00	100.00	100.00

Table 5.19.: Iris PAD performance reported in terms of TDR (%) at 0.2% BPCER on NDCLD-2013 and NDCLD-2015 databases. The best performance is in bold.

PAD Algorithms	Cogent	Vista
Textural Features [232]	55.53	87.06
WLBP [255]	65.40	66.91
LBP+SVM [102]	77.46	76.01
LBP+PHOG+SVM [12]	75.80	74.45
mLBP [239]	80.87	93.91
ResNet18 [107]	85.15	80.97
VGG [207]	90.40	94.82
MVANet [101]	94.90	95.11
DenseNet	99.37	**100.00**
PBS	99.62	**100.00**
A-PBS	**99.70**	**100.00**

Table 5.20.: Iris PAD performance in terms of CCR (%) on IIITD-CLI database. The best performance is in bold.

TDR at 0.2% BPCER is reported in Table5.19. The experiments are performed on Cognet and Vista sensor subsets, respectively. As shown in Table 5.18, our PBS and A-PBS solutions outperform all hand-crafted and CNN-based methods by a large margin (99.79% on Cognet subset and 100.00% on Vista subset). The near-perfect classification performance obtained by DenseNet, PBS, and A-PBS reveals that despite the significant PAD improvements achieved by deep learning models, there is an urgent need for large-scale iris PAD databases to be built for future research and generalizability analysis.

Iris PAD in the visible spectrum In addition to results on NIR databases, we also report results on the visible-light-based PAVID database in Table 5.21. The experiments were demonstrated following the defined protocols in [190]. For example, the Nokia - iPhone setup refers to the training and testing data as bona fide videos captured using the Nokia phone and the attack videos captured by iPhone. Moreover, we provide the results under a grand test setup, where bona fide and attack data includes videos captured by Nokia and iPhone. The models trained under grand-test setup will be used for cross-spectrum experiments later. It can be observed in Table 5.21 that deep-learning-based methods, including our A-PBS, outperform all the previously reported results on the PAVID database, which are hand-crafted feature-based PAD solutions. The DenseNet, PBS, and A-PBS methods obtain the best performance with all error rates of 0.00%.

Video	Metric	IQM-SVM [92, 190]	LBP-SVM [173, 190]	BSIF-SVM [188, 190]	STFT [190]	DenseNet	PBS	A-PBS
Nokia - iPhone	APCER	4.50	4.51	10.81	4.46	0.00	0.00	0.00
	BPCER	76.92	3.84	2.56	1.28	0.00	0.00	0.00
	HTER	40.71	4.18	6.68	2.87	**0.00**	**0.00**	**0.00**
Nokia - Nokia	APCER	3.57	2.67	0.89	2.68	0.00	0.00	0.00
	BPCER	57.31	4.87	6.09	1.21	0.00	0.00	0.00
	HTER	30.44	3.77	3.49	1.95	**0.00**	**0.00**	**0.00**
iPhone - iPhone	APCER	11.60	0.89	9.82	1.78	0.00	0.00	0.00
	BPCER	57.31	4.87	6.09	1.21	0.00	0.00	0.00
	HTER	34.45	2.88	7.96	1.49	**0.00**	**0.00**	**0.00**
iPhone - Nokia	APCER	10.71	3.54	8.92	0.00	0.00	0.00	0.00
	BPCER	76.92	3.84	2.56	1.28	0.00	0.00	0.00
	HTER	43.81	3.69	5.74	0.64	**0.00**	**0.00**	**0.00**
Grand-test	APCER	-	-	-	-	0.00	0.00	0.00
	BPCER	-	-	-	-	0.00	0.00	0.00
	HTER	-	-	-	-	**0.00**	**0.00**	**0.00**

Table 5.21.: Iris PAD performance of our proposed methods and established solutions on PAVID database with a threshold of 0.5. The results are reported based on APCER (%), BPCER (%), and HTER (%). Nokia - iPhone refers that the bona fide video is captured by Nokia while the replayed attack video is captured by iphone, and vice versa. Grand Test refers that both, bona fide and reply, videos are captured by Nokia and iphone. The best performance (the lowest HTER value) is in bold.

Cross-spectrum evaluation results

Most studies [155, 238] have presented PAD algorithms and verified their performance on NIR-based database. However, the performance of visible-light iris PAD has been understudied, especially under the cross-spectrum scenario. Therefore, we used the visible-light-based PAVID [190] and the NIR-based LivDet-Iris 2017 [243] databases to explore the effect of PAD performance across different spectra. The first scenario is the VIS-NIR where the models trained under the PAVID grand-test setup (visible spectrum) were evaluated on the test subsets of the NIR databses (Clarkson, Notre Dame, and IIIT-WVU), respectively. This evaluation results are presented in Table 5.22 and the bold numbers indicate the best performance (lowest error rates). It can be seen that our PBS

Train database	PAVID											
Test database	Clarkson				Notre Dame				IIITD-WVU			
Metric	EER	HTER	APCER	BPCER	EER	HTER	APCER	BPCER	EER	HTER	APCER	BPCER
DenseNet	37.78	36.69	**45.97**	27.41	56.39	56.69	59.28	54.11	54.43	49.94	**9.40**	49.94
PBS	**30.43**	37.12	66.23	**8.01**	55.67	55.39	81.22	**29.56**	51.10	50.59	82.66	18.52
A-PBS	33.41	**33.57**	46.20	20.94	**53.11**	**53.83**	**40.33**	65.89	**26.32**	**26.13**	36.30	**15.95**

Table 5.22.: Iris PAD performance measured under cross-spectrum scenarios and reported in terms of EER (%) and HTER (%), APCER(%), and BPCER(%). APCER and BPCER is determined by a threshold of 0.5. The training subset of the grand-test on the visible-light-based PAVID database is used to train a model, and the testing subset of each database in the LivDet-Iris 2017 database is used for evaluation. The lowest error rate is in bold.

Train database	Clarkson				Notre Dame				IIITD-WVU			
Test database	PAVID											
Metric	EER	HTER	APCER	BPCER	EER	HTER	APCER	BPCER	EER	HTER	APCER	BPCER
DenseNet	6.04	13.53	23.94	3.13	57.49	61.40	95.30	27.50	**8.28**	8.07	**7.38**	8.75
PBS	4.47	**5.97**	**10.07**	1.88	56.38	57.94	**76.51**	39.38	13.43	14.15	17.67	10.63
A-PBS	**1.34**	12.98	25.95	**0.00**	**52.35**	**50.63**	100.00	**1.25**	8.63	**8.05**	11.63	**5.62**

Table 5.23.: Iris PAD performance measured under cross-spectrum scenarios and reported in terms of EER (%), HTER (%), APCER(%), and BPCER(%). APCER and BPCER is determined by a threshold of 0.5. The training subset of the NIR-based Clarkson, Notre Dame, and IIITD-WVU is used to train a model, and the testing subset of the grand-test on the PAVID database is used for evaluation. The lowest error rate is in bold.

and A-PBS outperform the trained from scratch DenseNet. However, all PAD methods do not generalize well on the Notre Dame database. One possible reason is that Notre Dame comprises only challenging textured lens attacks and no print/reply attacks. The PAVID database, used for training here, only include reply attacks. One must note that both reply and print attacks involve the recapture of an artificially presented iris sample, unlike lens attacks. This recapture process can introduce artifacts identifiable by the PAD algorithms. Table 5.23 presents the results tested on the PAVID databases by using respective models trained on LivDet-Iris 2017 databases (the case of NIR-VIS). Similar observation can be found in Table 5.23 that the model trained on Notre Dame can not generalize on the PAVID database, e.g., the lowest EER and HTER values are 52.34% and 50.63% obtained by our A-PBS solution. In contrast to the results on Notre Dame, the model trained on Clarkson and IIITD-WVU generalizes much better on the visible-light database. The lowest EERand HTER values are 1.34% achieved by A-PBS and 5.97% achieved by PBS methods, while DenseNet and A-PBS obtained similar error rates on IIITD-WVU. Moreover, we illustrate the PAD score distribution with the fisher discriminant ratio [57, 171], which measures the separability, for further analysis. Figure 5.22 and 5.23 presents the results of case NIR-VIS and VIR-NIR, respectively. The PAD score distributions of the NIR-VIS case in Figure 5.22 show that models trained on Notre Dame perform worse than those trained on Clarkson and IIIT-WVU (bona fide and attack scores almost completely overlap). Moreover, the model trained on PAVID also obtained the largest overlapping and the smallest FDR value in Figure 5.23. One possible reason is the insufficient training data in Notre Dame (1,200 training data). However, the main reason might relate to the type of attacks and the lack of the recapturing process in the lens attacks, as mentioned earlier. This is also verified by the quantitative results in Table 5.22 and 5.23 (the APCER values are between 40.33% to 100.00%).

Visualization and Explainability

PBS is expected to learn more discriminative features by supervising each pixel/patch in comparison with binary supervised DenseNet. Subsequently, the A-PBS model, an extended model of PBS, is hypothesized to automatically locate the important regions that carry the features most useful for making an accurate iris PAD decision. To further verify and explain these assumptions, Score-Weighted Class Activation Mapping (Score-CAM) [226] is used to generate the visualizations for randomly chosen bona fide and attack iris images (these images belong to the same identity) under intra-database and cross-spectrum scenarios as shown in Figure 5.24 and 5.25.

Figure 5.24 illustrates the score-CAM results on the PAD samples in the test subset of IIITD-WVU. We adopted models trained on the training subset of IIITD-WVU (PAD) and models trained on the training subset of PAVID (visible-light) to generate score-CAMs, respectively. As shown in Figure 5.24, it is clear that PBS and A-PBS models pay more attention to the iris region than DenseNet in both intra-database and cross-spectrum cases. The DenseNet model seems to lose some information due to binary supervision. Similar observations can be found in Figure 5.25, where the PAD and visible-light models were tested on the visible images in the test subset of the PAVID database. In the visible intra-database case, DenseNet gained more attention on the eye region of visible-light images than of PAD images in Figure 5.24. Moreover, in the cross-spectrum case in Figure 5.25, the use of the attention module (A-PBS) has enabled the model to keep focusing on the iris area, while DenseNet and PBS lost some attention, especially on the attack samples. In general, the observations in Figures 5.24 and 5.25 are consistent with the quantitative results in Table 5.23 and 5.22 that the training on visible-light and test on PAD data (VIS-PAD) is more challenging than the training on PAD and test on visible-light data (PAD-VIS) in our case. It might be caused by: 1) The perceived image quality of visible data in the PAVID database are relatively lower than PAD images (see samples in Figure 5.20). 2) Some of the video frames in the PAVID database have an eye-blinking process, and thus some iris information (regions) will be hidden by eyelids and eyelashes. 3). While the used visible data (PAVID) contains only recaptured attacks (reply attacks), the PAD data contains both recaptured attacks (print attacks) and lens attacks, which makes it more difficult for a PAD trained on the visible images to perform properly on PAD attacks in our experiments.

5.5.4. Discussion

This section focused on the iris PAD performance in the PAD and visible domain, including challenging cross-database and cross-spectrum cases. The experiments were conducted using the proposed attention-based pixel-wise binary supervision (A-PBS) method [79] for iris PAD. A-PBS solution [79] aimed to capture the fine-grained pixel/patch-level cues and utilize regions that contribute the most to an accurate PAD decision by utilizing an attention mechanism. The extensive experiments were performed on six publicly available iris PAD databases in the PAD spectrum (including LivDet-Iris 2017 competition databases) and one database in the visible spectrum. By observing intra-database and intra-spectrum experimental results, several conclusion can be driven that 1) The results reported on respective attack types indicated that textured contact lens attack is more challenging to detect correctly than printouts attack. 2) cross-PA and cross-database are still challenging (EER values are over 20% in most cases). Furthermore, to our knowledge, this chapter is the first work to perform and analyze experiments under the cross-spectrum scenario. The experimental results showed that models trained on the visible spectrum do not generalize well on PAD data. It might also be caused by the limited visible data and its attack mechanism. In general, the A-PBS solution presents a superior PAD performance and high generalizability in the PAD and visible captured images, cross-database experiments, as well as cross-spectrum PAD deployments. The A-PBS [79] also showed to focus the attention of the PAD models towards the iris region when compared to more traditional solutions.

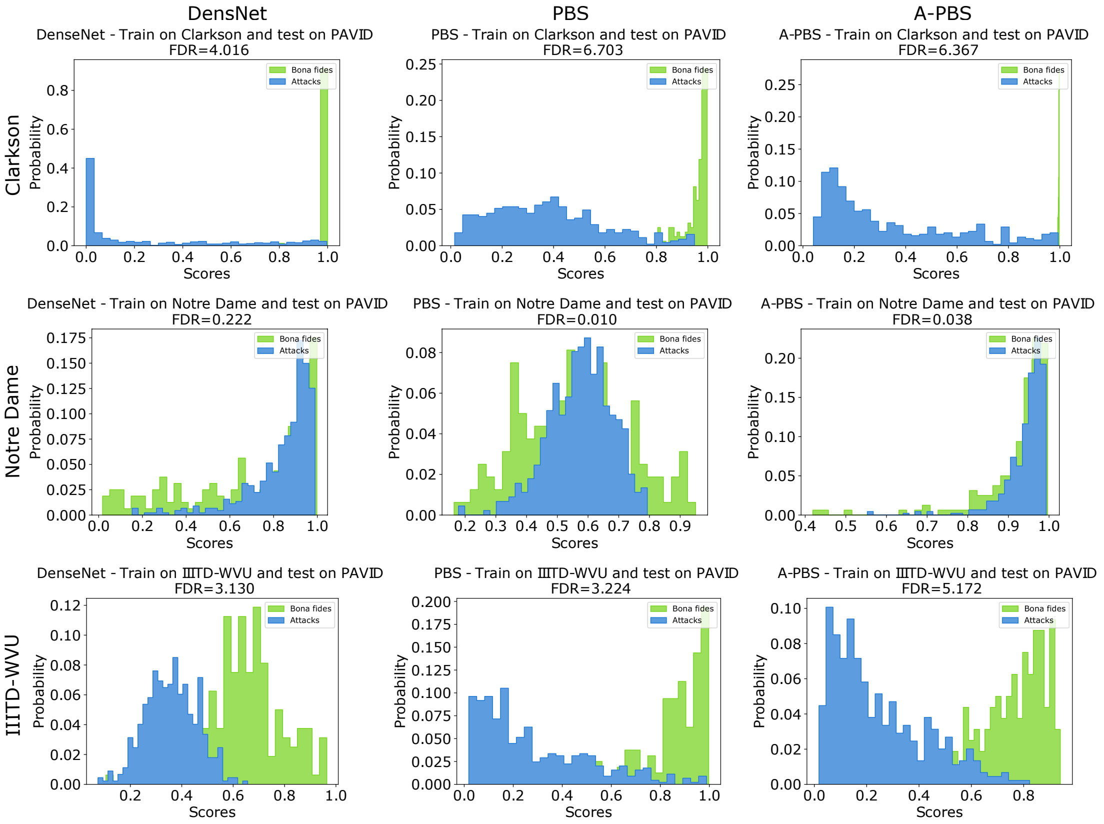


Figure 5.22.: PAD score distribution of bona fide (green) and PAs (blue) under cross-spectrum scenario (NIR-VIS). The models trained on the training subset of Clarkson (top), Notre Dame (middle), and IIITD-WVU (bottom) databases are used to evaluate on the test subset of PAVID database. and the histograms from left to right are produced by DenseNet, PBS, and A-PBS, respectively. The larger separability (measured by Fisher Discriminant Ratio (FDR) [57, 171]) and smaller overlap indicate higher classification performance.

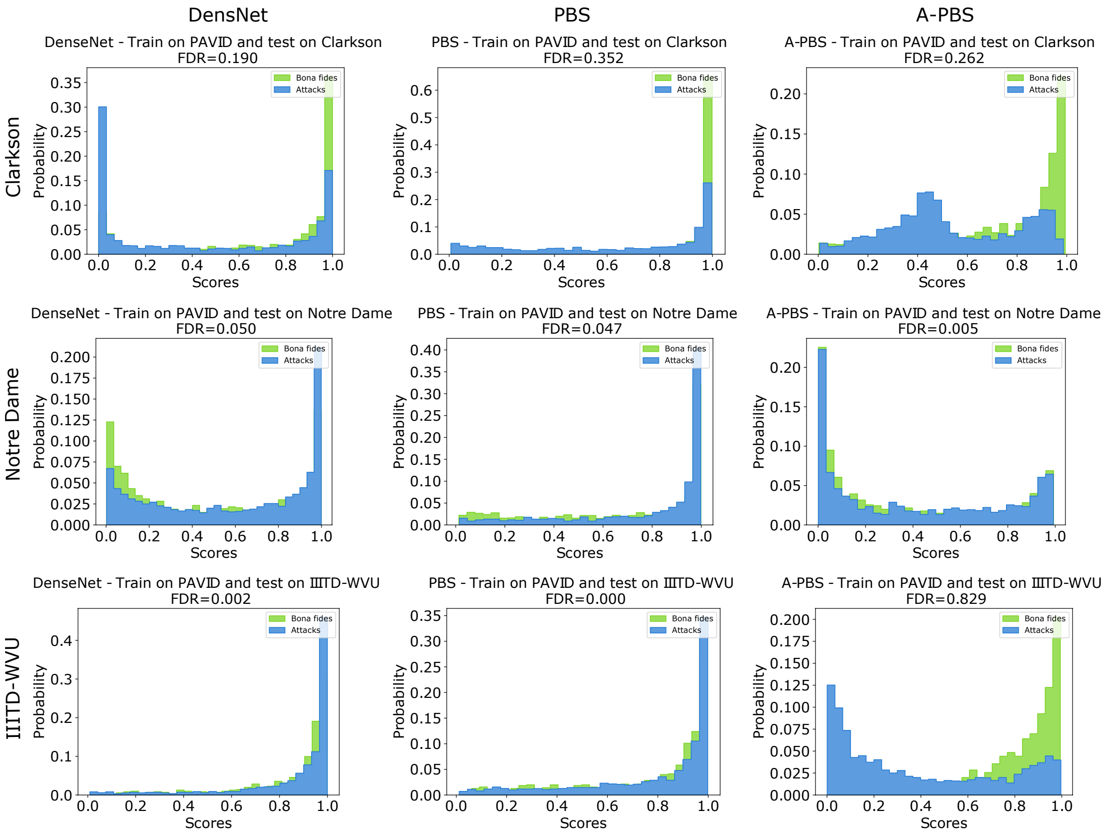


Figure 5.23.: PAD score distribution of bona fide (green) and PAs (blue) under cross-spectrum scenario (VIS-PAD). The model trained on PAVID database is used to test on the test subset of Clarkson, Notre Dame, and IIITD-WVU, respectively. The histogram top to bottom are test results on Clarkson, Notre Dame and IIITD-WVU databases, and the histograms from left to right are produced by DenseNet, PBS, and A-PBS, respectively. The larger separability (measured by Fisher Discriminant Ratio (FDR) [57, 171]) and smaller overlap indicate higher classification performance.

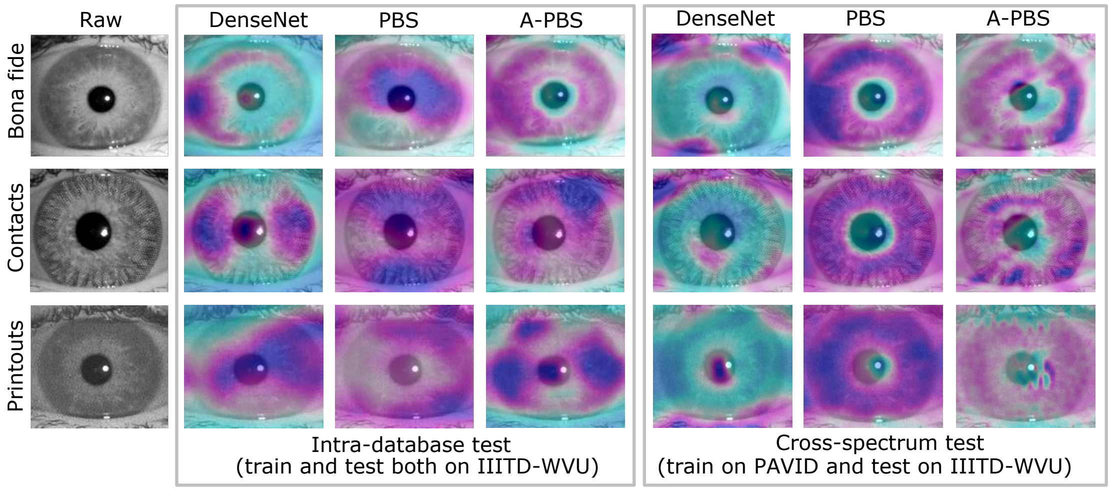


Figure 5.24.: Score-CAM visualizations for bona fide and attack samples in the IIITD-WVU test set under intra-database and cross-spectrum (VIS-PAD) scenarios. The darker the color of the region, the higher the attention on this area. The column from left to right refers to the raw samples, maps produced by DenseNet, PBS, and A-PBS model under two cases, respectively. The row from top to bottom refers to bona fide samples, textured contact lens, and printouts attack. PBS and A-PBS models pay more attention on iris region than DenseNet in both cases.

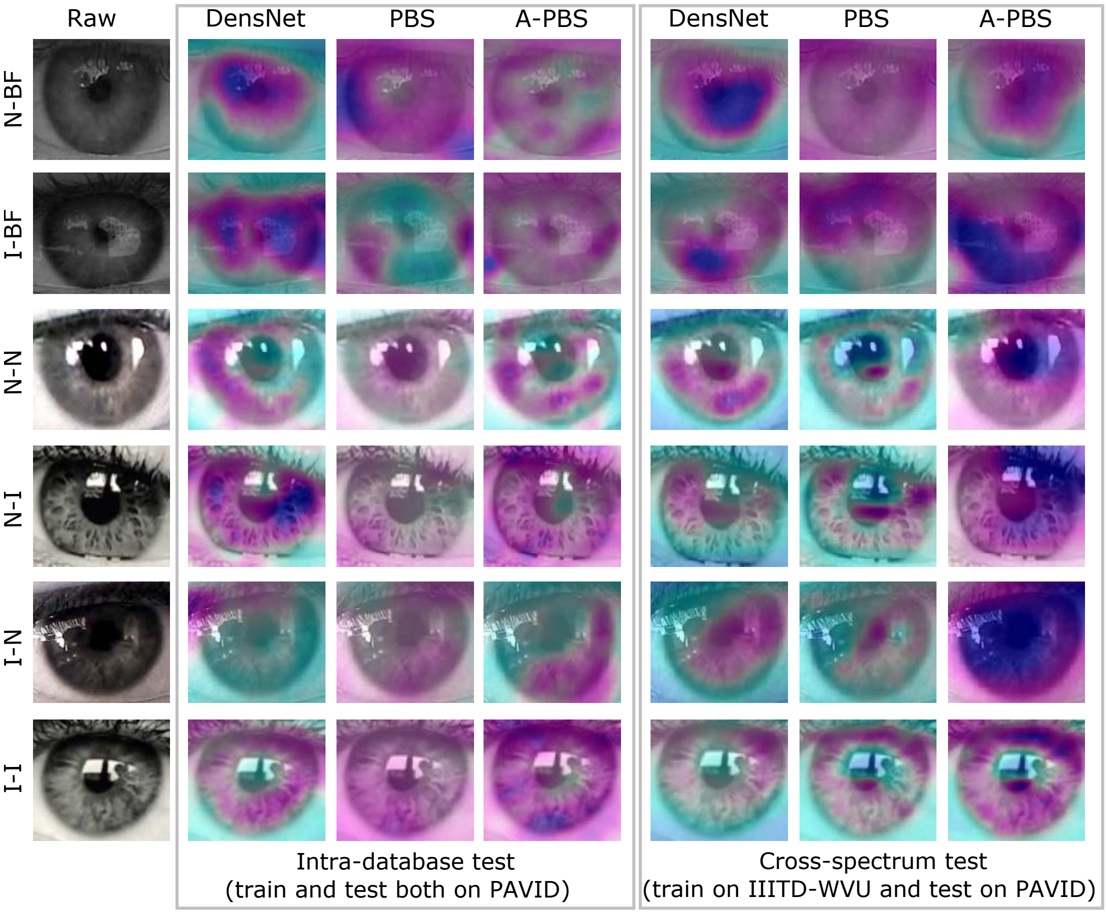


Figure 5.25.: Score-CAM visualizations for bona fide and attack samples in the PAVID test set under intra-database and cross-spectrum (PAD-VIS) scenarios. The darker the color of the region, the higher the attention on this area. The column from left to right refers to the raw samples, maps produced by DenseNet, PBS, and A-PBS model in two cases, respectively. N and I refer to Nokia and iPhone, respectively. The first two rows refer to the bona fide sample captured by Nokia and iPhone. The following four rows present the replay attack cases: Nokia-Nokia, Nokia-iPhone, iPhone-Nokia, and iPhone-iPhone. It is clear that the A-PBS model with an additional attention module is able to preserve relatively more attention on the iris region, especially for attack samples.

5.6. Summary

This chapter focused on the generalizability of PAD problem by presenting one face PAD solution and two iris PAD solutions. At first, this chapter presented a novel face PAD method by intelligently leveraging the information from the frequency domain to enhance the generalizability of face PAD. Later, the spatially aware iris regions were studied to enhance the performance and generalizability of the iris PAD. Finally, the attention mechanism-based iris PAD model was discussed under the cross-domain evaluation, including the cross-spectrum scenario.

Despite the great performance achieved by the hand-crafted and deep-learning-based face PAD methods in intra-dataset evaluations, the performance drops when dealing with unseen scenarios. The first part of this chapter proposed a dual-stream CNN framework. One stream adapted four learnable frequency filters to intelligently learn features in the frequency domain, which are less influenced by variations in sensors/illuminations. The other stream leveraged the RGB images to complement the features of the frequency domain. Moreover, a hierarchical attention module was integrated to fuse the information from the two streams at different stages by considering the nature of deep features in different layers of the CNN. The proposed method was evaluated in the intra-dataset and cross-dataset setups, and the results demonstrated that the proposed approach enhances the generalizability in most experimental setups in comparison to SOTA, including the methods designed explicitly for domain adaption/shift problems, as a response to *RQ6*. Furthermore, the design of the proposed PAD solution was proved successfully by a step-wise ablation study that involved the presented learnable frequency decomposition, the hierarchical attention module design, and the used loss function.

Existing iris PAD solutions are not deployable in mobile devices due to the large model size and lack the analysis of the generalizability under varied captured environments, unknown sensors or unknown attacks. To address these issues, in the second part of this chapter, the benefit of the spatially aware iris region was studied. Section 5.4 proposed a micro-stripe analyses solution to detect iris PAs. The proposed MSA method focused on the differences between bona fide and attack in the image dynamics around the iris/sclera boundary area. To achieve that, multiple thin micro-stripes of the normalized segmentation were extracted to provide more samples and less sample dimensionality, resulting in a better fitted model. The decision of multiple overlapping stripes was fused by a simple majority vote to build the final detection decision. In addition, the assumption that the iris/sclera boundary contains relatively more significant information for a PAD decision, even in the cases of unknown databases and unknown attacks, was supported by the illustrated EER heatmaps based on different pupil-centred rings. The lowest EER values commonly occurred in the fifth to the eighth stripe areas, which are those around the iris/sclera boundary. An in-depth experimental evaluation of this MSA method revealed a superior performance in three databases, in comparison with SOTA algorithms and baselines. Additionally, extensive cross-database and cross-attack (unknown-attack) detection evaluation experiments were demonstrated to explore the generalizability of the proposed method, texture-based method, and neural network-based methods. The results indicated that the MSA approach has better generalizability compared to other baselines in most experiments. Section 5.4 provided an answer to *RQ7* by studying the significance of different pupil-centred eye areas in iris PAD decisions using MSA method, and proving its improved generalizability under different experimental settings.

The third part of this chapter focused on the iris PAD performance in the NIR and visible domain, including challenging cross-database and cross-spectrum cases. First, it introduced a novel attention-based deep pixel-wise binary supervision method, named A-PBS, for iris PAD. A-PBS solution aimed to capture the fine-grained pixel/patch-level cues and utilize regions that contribute the most to an accurate PAD decision by utilizing an attention mechanism. The A-PBS showed to focus the attention of the PAD models towards the iris region when compared to more traditional solutions. The extensive experiments performed on six iris PAD databases in the NIR spectrum and one database in the visible spectrum verified the effectiveness and robustness of the presented A-PBS methods under scenarios with unknown attacks, sensors, and databases. Furthermore, this work performed and analyzed experiments under the cross-spectrum scenario by using A-PBS method. The experimental results showed that models trained on the visible spectrum do not generalize well on NIR data and the models trained on the NIR data generalized relatively better on visible data compared to the aforementioned cases. This could be attributed to the limited visible data and its attack mechanism. In general, the A-PBS solution presented a superior PAD performance and high generalizability in the NIR and visible captured images, cross-database experiments, as well as cross-spectrum PAD deployments, as an answer to *RQ8*.

Previous chapters assessed the fairness of PAD systems, discussed the emerging masked face PAD challenge, presented a set of solutions to enhance the generalizability of face and iris PAD. The next chapter will conclude the thesis and provide a brief outlook on future work.

6. Conclusion and Future Work

The previous chapters (3, 4 and 5) provided detailed responses to the research questions presented in Chapter 1. In the following, a summarized contribution of this thesis and an outlook for future research are provided.

6.1. Conclusion

The aim of presentation attack detection is to mitigate the vulnerability of biometric recognition algorithms to these attacks, thus enabling a wider deployment of biometric technology. This thesis presented advances in PAD by addressing three under-researched and emerging challenges of PAD. The contributions in this thesis were motivated by these three challenges identified in Chapter 1 and targeted the research questions presented in this thesis. The three main targeted challenges were fairness in face PAD, the emerging challenge of masked face PAD, and the generalizability of PAD.

Fairness in face PAD The first part of this thesis (Chapter 3) aims to analyse and enhance the fairness of face PAD, which itself can be viewed as a PAD generalizability issue. FR performance has been shown to be *unfair* to certain demographic and non-demographic groups. However, the fairness of face PAD is an understudied issue, mainly due to the lack of appropriately annotated data. To enable answering *RQ1*, a combined attribute annotated PAD dataset, CAAD-PAD, that combines several well-known PAD datasets were presented, in which seven human-annotated attribute labels were provided, covering both demographic and non-demographic attributes. To represent both the PAD fairness and the absolute PAD performance simultaneously, a novel metric, ABF, was introduced. Toward this end, *RQ1* was answered by comprehensively analysing the fairness of a set of face PADs and its relation to the nature of training data and the ODTA on different data groups on the presented CAAD-PAD dataset. Extensive experiments on CAAD-PAD showed that the nature of the training data and ODTA induced unfairness on gender, occlusion, and other attribute groups. Based on these analyses, a data augmentation method, FairSWAP, was proposed, as a response to *RQ2*. FairSWAP aimed to disrupt the identity/semantic information and guide models to mine attack clues rather than attribute-related information. Detailed experimental results demonstrated that FairSWAP generally enhanced both the PAD performance and fairness.

The emerging challenge of masked face PAD The second part of this thesis (Chapter 4) targeted the emerging and unexpected challenge posed by wearing a facial masked during the COVID-19 pandemic on FR and PAD. Since existing face PAD databases only contain attacks in which subjects were not wearing face masks. The relationship between the performance of PAD techniques and masked faces is uncertain and the vulnerability of FR systems to masked face attacks remains unclear. To overcome such gaps and enable the study, CRMA dataset was presented, including the conventional unmasked bona fides and attacks, novel bona fide and attacks with faces wearing masks, and attacks with real masks placed on spoof faces. It consists of 13,113 high-resolution videos and has a large diversity in capture sensors, displays, and capture scales. Moreover, three experimental protocols were designed to study the effect of wearing a mask on the PAD algorithms.

As a response to *RQ3*, this thesis presented a thorough analysis of the vulnerability of FR systems to such masked faces. The results indicate that FR systems are vulnerable to both masked and unmasked attacks. For example, when the reference images and system threshold are based on unmasked bona fide faces (BM0-BM0), the IAPMR values for unmasked attacks (AM0), masked attacks (AM1), and attacks covered by a real mask (AM2) are 98.40%, 81.60%, and 97.10%, respectively. This leads to the interesting observation that all the investigated FR systems are more vulnerable to attacks where real masks are placed on attacks (AM2) than attacks of masked faces (AM1).

Additionally, this thesis conducted extensive experiments by following designed protocols, providing an answer to *RQ4*. The experimental results indicated that PAD algorithms have a high possibility of detecting masked bona fide samples as attackers (median BPCER value for BM1 in protocol-1 is 48.25%). Even if PAD solutions have seen the masked bona fide data during training, PAD algorithms still perform worse on masked bona fide samples compared to unmasked bona fide ones. Furthermore, PAD solutions trained on masked face attacks (AM1) do not generalize well on attacks covered by a real mask (AM2).

Lastly, a response to *RQ5* was drawn by presenting a solution to target the masked PAs, especially partially covered attacks, by proposing both the partial attack supervision and the regional weighted inference. The aim of partial attack supervision was to guide models to better convergence while training. Meanwhile, regional weighting improved the generalizability of PAD in the inference phase. The detailed ablation study showed the consistent benefits of both components, separately and jointly, in a single solution. The presented PAL-RW based models outperformed other established PAD methods when dealing with the possibility of masked faces in PAD decisions. Moreover, as PAL-RW focused on refining training labels and post-processing of PAD predictions, it thus could be easily incorporated into any neural network architecture.

Generalizability of PAD algorithms The third part of this thesis (Chapter 5) aims at enhancing the generalizability of PAD by presenting one face and two iris PAD solutions.

As a response to *RQ6*, this thesis proposed a novel dual-stream face PAD framework by intelligently exploiting information from the frequency domain to enhance the generalizability of face PAD. One of the streams adapted four learnable frequency filters to learn features in the frequency domain in an optimized manner, which is less influenced by variations in sensors/illuminations. The other stream leveraged RGB images to complement the features in the frequency domain. Moreover, a hierarchical attention module was integrated to fuse the information from two streams at different stages by considering the nature of deep features in different layers of CNNs. Through extensive intra-dataset and cross-dataset evaluations and comparisons to the recent SOTA algorithms, including methods explicitly for domain adaption/shift problems, the experimental results proved the superiority of the proposed method over recent PAD algorithms, especially under most cross-dataset scenarios. The design of the proposed PAD solution was also successfully demonstrated by a step-wise ablation study involving the presented learnable frequency decomposition, the hierarchical attention module design, and the used loss function.

In the aspect of iris PAD generalizability, this thesis responded to *RQ7* by studying the benefit of spatially aware iris and iris/sclera region. This is performed by proposing the micro-stripe analyses solution, MSA, to detect iris PAs. The proposed MSA method focused on the differences between bona fide and attack in the image dynamics around the iris/sclera boundary areas. The illustrated EER heatmaps based on different pupil-centred rings supported our assumption that the iris/sclera boundary contains relatively more significant information for a PAD decision, even in the cases of unknown databases and unknown attacks. As a result, multiple thin micro-stripes of the normalized segmentation were extracted to provide more samples and less sample dimensionality, resulting in a better-fitted model. The final decision was determined by a simple majority vote of multiple overlapping stripes. An in-depth experimental evaluation of MSA method revealed a superior performance on mainstream iris PAD datasets, in comparison with SOTA algorithms and baselines. Extensive cross-database and cross-attack (unknown-attack) evaluation experiments were conducted and proved the enhanced generalizability of iris PAD obtained by MSA.

Following the need to build automatic decisions based on correctly localized clues, the third part of Chapter 5 introduced a novel attention-based deep pixel-wise binary supervision method, named A-PBS. The A-PBS solution aimed to capture the fine-grained pixel/patch-level clues and automatically locate regions that contribute the most to an accurate PAD decision by using an attention mechanism. To demonstrate the effectiveness and robustness of the presented A-PBS, a set of extensive experiments were conducted on six NIR iris PAD databases and one visible spectrum database by following the protocols for the unknown attack, sensor, database, and spectrum scenarios. The results reported in Chapter 5 showed that models trained on visible spectrum did not generalize well on NIR data, while models trained on NIR data generalized relatively better on visible data compared to the aforementioned cases. This could be attributed to limited visible data and its attack mechanism. Overall, the A-PBS solution obtained a superior PAD performance and boosted generalizability on both NIR and visible spectrum intra-dataset scenarios, cross-database experiments, as well as cross-spectrum PAD deployments, providing an answer to *RQ8*.

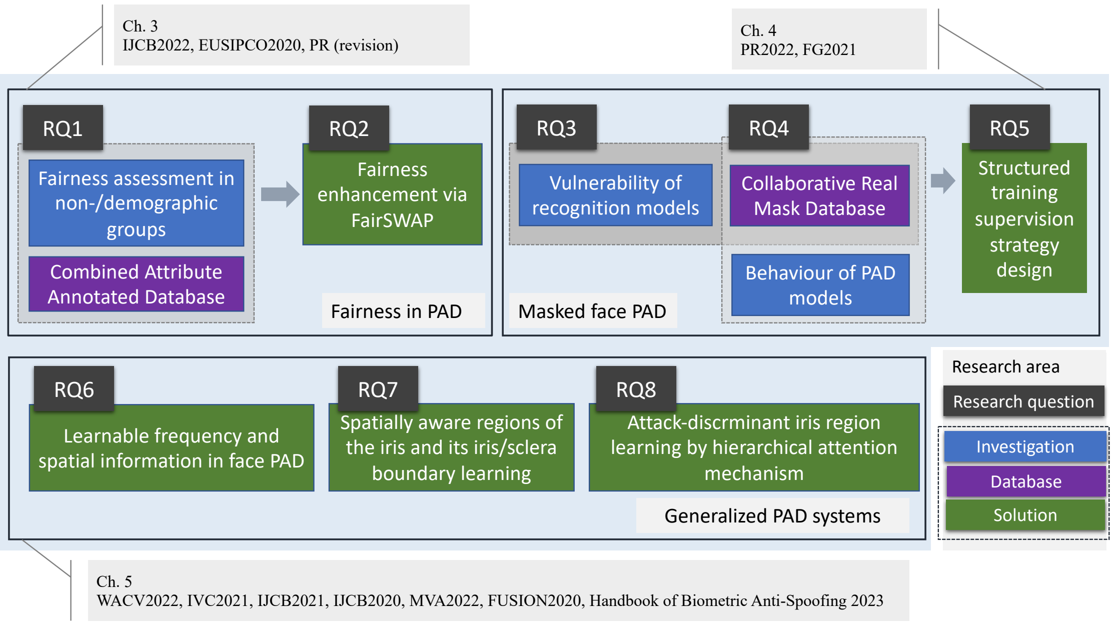

Ch. 3
IJCB2022, EUSIPCO2020, PR (revision)
Ch. 4
PR2022, FG2021
RQ1
Fairness assessment in non-/demographic groups
Combined Attribute Annotated Database
RQ2
Fairness enhancement via FairSWAP
Fairness in PAD
RQ3
Vulnerability of recognition models
RQ4
Collaborative Real Mask Database
Behaviour of PAD models
Masked face PAD
RQ5
Structured training supervision strategy design
RQ6
Learnable frequency and spatial information in face PAD
RQ7
Spatially aware regions of the iris and its iris/sclera boundary learning
RQ8
Attack-discrminant iris region learning by hierarchical attention mechanism
Generalized PAD systems
Research area
Research question
Investigation
Database
Solution
Ch. 5
WACV2022, IVC2021, IJCB2021, IJCB2020, MVA2022, FUSION2020, Handbook of Biometric Anti-Spoofing 2023

Figure 6.1.: A summary of contributions in relation with the research questions posed in this thesis, the chapters responding to those questions, and the publications building to these chapters.

Summary To sum up, this thesis focused on boosting the generalizability and fairness of PAD by presenting a set of contributions driven by a set of research questions listed in Chapter 1. Figure 6.1 links these research questions to the corresponding chapters along with the related publications.

In response to the *RQ1: Are PAD systems fair? And do the lack of balance in PAD training data and the methodology of choosing PAD decision threshold affect this fairness?*, this thesis confirmed that PAD systems are unfair to certain groups and fairness of face PADs is affected by the nature of training data and the ODTA. This answer is confirmed by comprehensively analyzing the fairness of face PADs and their relation to both effect factors through extensive experiments on the presented CAAD-PAD dataset. Knowing the outcome of *RQ1*, this thesis proposed a simple solution, FairSWAP, serving as a data augmentation technique, to enhance the fairness of investigated PAD solutions, providing a confirmed answer to *RQ2: Is the fairness of PAD solutions enhanced when controllably augmenting the training data so that different data groups will possess specific properties of other groups?*.

Towards the masked face PAD challenge, this thesis first presented CRMA dataset to enable answering *RQ3*, *RQ4*, and *RQ5*. In response to *RQ3: Is the vulnerability of FR systems to PAs affected by wearing a mask?*, this thesis confirms that the vulnerability of FR systems differed when faced with masked faces. Compared to unmasked face attacks, FR systems are less vulnerable to both types of masked faces. Additionally, they were more vulnerable to attacks with real face masks (AM2) than masked face attacks (AM1). These conclusions were drawn by presenting a thorough analysis of the vulnerability of FR systems to novel masked faces in CRMA. The answer to *RQ4: Is the behaviour of existing PADs affected by wearing face masks, whether on bona fide or attack faces?* confirmed that the masked bona fide and attacks caused a significant PAD performance degradation by conducting a set of extensive experiments. Knowing such outcomes, this thesis further proposed a solution to target this issue, especially partially covered attacks. This solution is based on two modules, the partial attack supervision and the regional weighted inference. The improved PAD performance by the proposed solution provided a confirmed answer to *RQ5: Can the PAD performance, especially on masked samples, be enhanced by designing a structured PAD training supervision strategy that takes the possibility of a masked face into account?*.

Towards enhancing the generalizability of PAD algorithms, this thesis presented three contributions. First, to answer *RQ6: Can the generalizability of face PAD be enhanced by learning to include information from the frequency domain in an optimized manner?*, this thesis successfully designed a face PAD framework that adapted learnable frequency filters to intelligently learn features in the frequency domain, along with features from the spatial domain, proving to enhance the PAD generalizability. In response to *RQ7: Can analyzing spatially aware regions of the iris and its iris/sclera boundary enhance the generalizability of* *iris PAD?*, this thesis confirmed that by studying the benefits of spatially aware iris and its iris/sclera region and successfully introducing a micro-stripe analyses solution to detect iris PAs under both intra-dataset and cross-dataset evaluations. Furthermore, this thesis confirmed *RQ8: Can networks be automatically guided to focus on the attack-discriminant iris region during the PAD training? If so, does this enhance the iris PAD generalization?* by introducing a novel attention-based deep pixel-wise binary supervision method for iris PAD that successfully demonstrated the enhanced generalized performance under intra-dataset, cross-dataset, and cross-spectrum scenarios.

6.2. Future Work

As PAD plays an important role in protecting biometric systems from PAs, these PAD systems are even more demanding, and thus prompting several future research directions based on the contributions of this thesis. These research directions can be summarized as follows:

Efficient PAD solutions Due to the significant importance of PAD in mitigating the vulnerability of recognition algorithms, most researchers are competing to present solutions to boost the performance by employing overparameterized deep learning networks with high computational cost [108, 117, 207]. Despite the success of such deep learning models, the deployment of these overparameterized models on many use-cases is limited by the computational capability. This challenge has attracted increasing attention in biometric recognition tasks [18, 40, 175], but much less attention in PAD. Considering the use-cases that are extremely limited by computational resources like mobile devices, designing an accurate PAD solution with low complexity is crucial for deployability. This thesis has attempted to select MixFaceNet [18] as one of the backbones for building PAL-RW solution in Chapter 4 by considering its lower computational complexity and high accuracy in face verification and identification tasks. PAL-RW$_{MixFaceNet}$ obtained the best overall PAD performance, indicating the efficiency of MixFaceNet [18] and rationalizes the choice of this efficient backbone based on the different sizes of convolutional kernels. In addition to the manual design of efficent architectures, the core metrics of a given model can be improved by automatically designing some of its components. An example of that is using neural architecture search (NAS) [26, 165] to automate the design of neural architectures. In such a solutions, the architecture itself is finetuned and the automated search helps find a model that optimizes both loss/accuracy and other metrics such as model latency/size. Yu *et al.* [250] proposed a NAS-based PAD solution to discover suitable task-aware networks. However, this solution focused on studying the search spaces for performance enhancement, i.e. still possessing high computational cost. In addition to the efficient architecture design (manually or automatically), model compression techniques, including parameter pruning and model quantization, can be employed to reduce the required computational cost of deep learning models. Overall, efficient PAD solutions raise the need for considering and reporting the cost of model development, training, and deployment.

Standardization of fairness Fairness in AI systems has attracted increasing attention and gained significant importance in building such systems. Current fairness research concentrates on the definition [123, 177, 224], metric [64, 98, 105], assessment, and enhancement [8, 64, 69, 208, 216, 218, 221, 252] of fairness. However, up to now, fairness is still a broad concept without a unifying standard definition. A recently released AI standardization document, ISO/IEC TR 24027:2021 [128], used the terms Fairness and Bias to represent AI bias and discussed the possibility of assessing fairness in AI with a fairness measure. This measure is currently being utilized in the biometric research community. Additionally, there are efforts to standardize the definition and measurement of fairness in biometric solutions with a standard currently under development [129], as fairness measures from AI are seen as lacking consideration for social, legal, or cognitive aspects of fairness. As suggested in [202], bias may exist at every stage of building AI systems, data collection and pre-processing, model/algorithm development, and validation/test step. This thesis only studied the fairness of PAD systems related to the nature of training data at the data collection stage and ODTA at the deployment stage. The problem formulation, model selection, or optimization objective may also affect the fairness at the model development stage. Such aspects are worth investigating in the future to improve the model design. Moreover, the existing fairness in biometric systems is mostly measured based on the differential performance independent of any thresholds or differential outcome by a group-specific threshold [4, 69, 81]. This thesis thus adapted a recently defined fairness metric based on a decision threshold computed from all groups, and proposed ABF to associate fairness with the absolute PAD performance. However, as investigated in Chapter 3, the fairness of face PADs was influenced by ODTAs. Therefore, it is critical to define standard metrics for fairness over varied use cases. Overall, the standardization of fairness will subsequently support fairness studies of data-driven and automated decision-making biometric systems and help evaluate and develop fairness enhancement algorithms.

PAD datasets Compared to datasets used in general computer vision tasks, such as ImageNet for object classification and detection, biometric data is relatively very limited due to the time and effort required to collect and manage it, along with privacy concerns associated with acquiring such datasets. As aforementioned in Chapter 3, most PAD datasets lack detailed annotations on attributes, and thus fairness in PAD has been poorly investigated. Besides, the presented CRMA dataset in Chapter 4 is also limited in terms of the number of subjects and the types of PAI. Moreover, due to the increasing ethical and legal problems, several biometric datasets are recently retracted by their creators, such as VGGFace2 [28], MS-Celeb-1M [100], and MegaFace [145] for face recognition, and Warsaw dataset for iris PAD and SiW for face PAD. To broaden the scope of the research field related to PAD, building large-scale PAD datasets is valuable. In addition to collecting authentic data, a major alternative can be building data based on synthetically generate data. Synthetic data has received increasing attention in the biometric field and successfully employed in face recognition [22, 187] and face morphing attack detection [55, 120]. Overall, large-scale PAD data with rich attribute annotations promote the diversity and depth of future research, including generalizability and fairness in PAD.

Domain generalization Despite the enhanced face and iris PAD generalizability by presented algorithms in Chapter 5, PAD is still an open problem in practice due to the large uncertainty gaps caused by the unknown nature of variations in the attacks. PAD is also a self-evolving problem, where PAs and detection algorithms evolve iteratively. One limitation of all the presented algorithms in Chapter 5 is that they are supervised learning, highly relying on the known presentations. Several unsupervised learning techniques could be adapted to learn a generalized feature representation. For example, recent task-independent self-supervised feature learning techniques [41, 106] showed promising results on general downstream computer vision tasks. Moreover, to alleviate the PAD domain bias problem, one possible direction is curriculum learning technique [103, 229], which imitates the learning strategy in human curricula, training models in a meaningful order (easy-to-hard etc.). The curriculum learning techniques have been proven to be successfully employed in many machine learning fields in a wide range of tasks [119, 211]. The enhanced performance on such tasks suggests that curriculum learning techniques can help the unsupervised PAD domain adaption research. Another limitation of current face PAD solutions is that many domain adaptation and domain generalization-based PAD methods rely on a priori knowledge, such as domain labels, and require access to the unlabeled target data, which are generally unavailable in real-world scenarios. Therefore, exploring the intrinsic causal mechanisms or performing domain adaptation/generalization on the embedding space could be potential future research directions to address this challenge.

A. Publications

The author published 13 scientific publications as a first author, contributed to 12 scientific publications.

A.1. Publications

1. Meiling Fang, Marco Huber, and Naser Damer. SynthASpoof: Developing Face Presentation Attack Detection Based on Privacy-friendly Synthetic Data. In *CVPR Workshops*, pages 1605–1616. IEEE, 2023

2. Meiling Fang, Fadi Boutros, and Naser Damer. Intra and cross-spectrum iris presentation attack detection in the nir and visible domains. In Sébastien Marcel, Julian Fierrez, and Nicholas Evans, editors, *Handbook of Biometric Anti-Spoofing: Presentation Attack Detection and Vulnerability Assessment*, pages 171–199, Singapore, 2023. Springer Nature Singapore

3. Meiling Fang, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Real masks and spoof faces: On the masked face presentation attack detection. *Pattern Recognition*, 123:108398, 2022

4. Meiling Fang, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Learnable multi-level frequency decomposition and hierarchical attention mechanism for generalized face presentation attack detection. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, Waikoloa, HI, USA, January 3-8, 2022*, pages 1131–1140. IEEE, 2022

5. Meiling Fang, Fadi Boutros, and Naser Damer. Unsupervised face morphing attack detection via self-paced anomaly detection. In *International IEEE Joint Conference on Biometrics, IJCB 2022, Abu Dhabi, United Arab Emirates, October 10-13, 2022*, pages 1–8. IEEE, 2022

6. Meiling Fang, Hamza Ali, Arjan Kuijper, and Naser Damer. PatchSwap: Boosting the Generalizability of Face Presentation Attack Detection by Identity-aware Patch Swapping. In *International IEEE Joint Conference on Biometrics, IJCB 2022, Abu Dhabi, United Arab Emirates, October 10-13, 2022*, pages 1–8. IEEE, 2022

7. Meiling Fang, Naser Damer, Fadi Boutros, Florian Kirchbuchner, and Arjan Kuijper. The overlapping effect and fusion protocols of data augmentation techniques in iris PAD. *Mach. Vis. Appl.*, 33(1):8, 2022

8. Meiling Fang, Naser Damer, Fadi Boutros, Florian Kirchbuchner, and Arjan Kuijper. Cross-database and cross-attack iris presentation attack detection using micro stripes analyses. *Image Vis. Comput.*, 105:104057, 2021

9. Meiling Fang, Fadi Boutros, Arjan Kuijper, and Naser Damer. Partial attack supervision and regional weighted inference for masked face presentation attack detection. In *16th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2021, Jodhpur, India, December 15-18, 2021*, pages 1–8. IEEE, 2021

10. Meiling Fang, Naser Damer, Fadi Boutros, Florian Kirchbuchner, and Arjan Kuijper. Iris presentation attack detection by attention-based and deep pixel-wise binary supervision network. In *International IEEE Joint Conference on Biometrics, IJCB 2021, Shenzhen, China, August 4-7, 2021*, pages 1–8. IEEE, 2021

11. Meiling Fang, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Demographic bias in presentation attack detection of iris recognition systems. In *28th European Signal Processing Conference, EUSIPCO 2020, Amsterdam, Netherlands, January 18-21, 2021*, pages 835–839. IEEE, 2020

12. Meiling Fang, Naser Damer, Fadi Boutros, Florian Kirchbuchner, and Arjan Kuijper. Deep learning multi-layer fusion for an accurate iris presentation attack detection. In *IEEE 23rd FUSION, online, July 6-9, 2020*, pages 1–8. IEEE, 2020

13. Meiling Fang, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Micro stripes analyses for iris presentation attack detection. In *2020 IEEE International Joint Conference on Biometrics, IJCB 2020, online, Sep.28 - Oct.1, 2020*, pages 1–10. IEEE, 2020

14. Fadi Boutros, Meiling Fang, Marcel Klemt, Biying Fu, and Naser Damer. CR-FIQA: face image quality assessment by learning sample relative classifiability. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, Canada, June 18-22, 2023*, pages 1–10. IEEE (to appear), 2023

15. Fadi Boutros, Marcel Klemt, Meiling Fang, Arjan Kuijper, and Naser Damer. Unsupervised face recognition using unlabeled synthetic data. In *FG*, pages 1–8. IEEE, 2023

16. Naser Damer, Meiling Fang, Patrick Siebke, Jan Niklas Kolf, Marco Hube, and Fadi Boutros. Mordiff: Recognition vulnerability and attack detectability of face morphing attacks created by diffusion autoencoders. In *11th IEEE International Workshop on Biometrics and Forensics, IWBF 2023, Barcelona, Spain, April 19-20, 2023*, pages 1–6. IEEE (to appear), 2023

17. Naser Damer, Fadi Boutros, Marius Süßmilch, Meiling Fang, Florian Kirchbuchner, and Arjan Kuijper. Masked face recognition: Human versus machine. *IET Biom.*, 7(1):15–26, 2022

18. Naser Damer, César Augusto Fontanillo López, Meiling Fang, Noémie Spiller, Minh Vu Pham, and Fadi Boutros. Privacy-friendly Synthetic Data for the Development of Face Morphing Attack Detectors. In *CVPR Workshops*, pages 1605–1616. IEEE, 2022

19. Fadi Boutros, Olga Kaehm, Meiling Fang, Florian Kirchbuchner, Naser Damer, and Arjan Kuijper. Low-resolution iris recognition via knowledge transfer. In *2022 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–5. IEEE, 2022

20. Fadi Boutros, Naser Damer, Meiling Fang, Florian Kirchbuchner, and Arjan Kuijper. Mixfacenets: Extremely efficient face recognition networks. In *International IEEE Joint Conference on Biometrics, IJCB 2021, Shenzhen, China, August 4-7, 2021*, pages 1–8. IEEE, 2021

21. Naser Damer, Kiran B. Raja, Marius Süßmilch, Sushma Venkatesh, Fadi Boutros, Meiling Fang, Florian Kirchbuchner, Raghavendra Ramachandra, and Arjan Kuijper. ReGenMorph: Visibly Realistic GAN Generated Face Morphing Attacks by Attack Re-generation. In *ISVC (1)*, volume 13017 of *Lecture Notes in Computer Science*, pages 251–264. Springer, 2021

22. Naser Damer, Noémie Spiller, Meiling Fang, Fadi Boutros, Florian Kirchbuchner, and Arjan Kuijper. PW-MAD: Pixel-Wise Supervision for Generalized Face Morphing Attack Detection. In *ISVC (1)*, volume 13017 of *Lecture Notes in Computer Science*, pages 291–304. Springer, 2021

23. Fadi Boutros, Naser Damer, Meiling Fang, Kiran B. Raja, Florian Kirchbuchner, and Arjan Kuijper. Compact models for periocular verification through knowledge

distillation. In *BIOSIG 2020 - Proceedings of the 19th International Conference of the Biometrics Special Interest Group, online, 16.-18. September 2020*, volume P-306 of *LNI*, pages 291–298. Gesellschaft für Informatik e.V., 2020

24. Sandip Purnapatra, Nic Smalt, Keivan Bahmani, Priyanka Das, David Yambay, Amir Mohammadi, Anjith George, Thirimachos Bourlai, Sébastien Marcel, Stephanie Schuckers, Meiling Fang, Naser Damer, Fadi Boutros, Arjan Kuijper, Alperen Kantarci, Basar Demir, Zafer Yildiz, Zabi Ghafoory, Hasan Dertli, Hazim Kemal Ekenel, Son Vu, Vassilis Christophides, Dashuang Liang, Guanghao Zhang, Zhanlong Hao, Junfu Liu, Yufeng Jin, Samo Liu, Samuel Huang, Salieri Kuei, Jag Mohan Singh, and Raghavendra Ramachandra. Face liveness detection competition (livdet-face) - 2021. In *IJCB*, pages 1–10. IEEE, 2021

25. Priyanka Das, Joseph McGrath, Zhaoyuan Fang, Aidan Boyd, Ganghee Jang, Amir Mohammadi, Sandip Purnapatra, David Yambay, Sébastien Marcel, Mateusz Trokielewicz, Piotr Maciejewicz, Kevin W. Bowyer, Adam Czajka, Stephanie Schuckers, Juan E. Tapia, Sebastian Gonzalez, Meiling Fang, Naser Damer, Fadi Boutros, Arjan Kuijper, Renu Sharma, Cunjian Chen, and Arun Ross. Iris liveness detection competition (livdet-iris) - the 2020 edition. In *IEEE IJCB, online, 28.09-01.10, 2020*, pages 1–9. IEEE, 2020

For an updated publication list, please check my DBLP and Google Scholar profiles:

- Goolge scholar: `https://scholar.google.com/citations?user=S_RV3L4AAAAJ&hl`
- Dblp:`https://dblp.org/pid/260/0316.html`

A.2. Submitted papers

1. Meiling Fang, Wufei Yang, Arjan Kuijper, Vitomir Struc, and Naser Damer. Fairness in face presentation attack detection. *CoRR*, abs/2112.06592, 2022 to *Pattern Recognition (revision)*

2. Meiling Fang and Naser Damer. Face presentation attack detection by excavating causal clues and adapting embedding statistics. *CoRR*, abs/, 2023 to *ICCV 2023 (submitted)*

A.3. Competitions

The author participated in two international competitions [61, 185].

1. The first place in image category and the second place in video category at Face Liveness Detection Competition (LivDet-Face) - 2021 [185]

2. The second place at Iris Liveness Detection Competition (LivDet-Iris) - 2020 [61]

B. Supervising thesis

The following bachelor and master theses were supervised by the author, and the results of these works were partially used as an input into this thesis.

1. Wufei Yang, Meiling Fang (supervisor) and Prof. Dr. Arjan Kuijper (supervisor). Bias Exploration and Mitigation in Face Presentation Attack Detection systems. Bachelor Thesis, TU Darmstadt, 2022.

2. Hamza Ali, Meiling Fang (supervisor) and Prof. Dr. Arjan Kuijper (supervisor). Intra-identity PatchSwap: On the Generalizability of Face Presentation Attack Detection. Master Thesis, TU Darmstadt, 2022.

3. Nour Eldin Alaa Badr, Meiling Fang (supervisor) and Prof. Dr.-Ing. Franz Kummert (supervisor). Momentum Contrast for Representative Face Presentation Attack Detection. Master Thesis, FH Bielefeld, 2022.

Bibliography

[1] *Handbook of Face Recognition, 2nd Edition*. Springer, 2011.

[2] ISO/IEC JTC 1/SC 27 Information security, cybersecurity and privacy protection. ISO/IEC 24745:2022 Information security, cybersecurity and privacy protection — Biometric information protection. International Organization for Standardization, 2022.

[3] Vítor Albiero, Krishnapriya K. S, Kushal Vangara, Kai Zhang, Michael C. King, and Kevin W. Bowyer. Analysis of gender inequality in face recognition accuracy. In *WACV Workshops*, pages 81–89. IEEE, 2020.

[4] Norah Alshareef, Xiaohong Yuan, Kaushik Roy, and Mustafa Atay. A study of gender bias in face presentation attack and its mitigation. *Future Internet*, 13(9):234, 2021.

[5] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[6] Sarah Baker, Amanda Hentz, Kevin Bowyer, and Patrick Flynn. Degradation of iris recognition performance due to non-cosmetic prescription contact lenses. *Computer Vision and Image Understanding*, 114:1030–1044, 09 2010.

[7] BBC News. Doctor 'used silicone fingers' to sign in for colleagues, 2013. `https://www.bbc.com/news/world-latin-america-21756709`.

[8] J. Ross Beveridge, Geof H. Givens, P. Jonathon Phillips, and Bruce A. Draper. Factors that influence algorithm performance in the face recognition grand challenge. *Comput. Vis. Image Underst.*, 113(6):750–762, 2009.

[9] Sushil Bhattacharjee, Amir Mohammadi, and Sébastien Marcel. Spoofing deep face recognition with custom silicone masks. In *9th IEEE International Conference on*

Biometrics Theory, Applications and Systems, BTAS 2018, Redondo Beach, CA, USA, October 22-25, 2018, pages 1–7. IEEE, 2018.

[10] ISO/IEC JTC 1/SC 37 biometrics. In information technology—biometric presentation attack detection—part 1: Frame-work. Iso, International Organization for Standardization, Geneva, Switzerland, 2016.

[11] Ruud Bolle, Sharath Pankanti, and Anil K. Jain. *Biometrics, Personal Identification in Networked Society: Personal Identification in Networked Society*. Kluwer Academic Publishers, USA, 1998.

[12] Anna Bosch, Andrew Zisserman, and Xavier Muñoz. Representing shape with a spatial pyramid kernel. In *CIVR 2007, Amst., The Netherlands, July 9-11, 2007*, pages 401–408. ACM, 2007.

[13] Zinelabdine Boulkenafet, Jukka Komulainen, Zahid Akhtar, Azeddine Benlamoudi, Djamel Samai, Salah Eddine Bekhouche, Abdelkrim Ouafi, Fadi Dornaika, Abdelmalik Taleb-Ahmed, Le Qin, Fei Peng, L. B. Zhang, Min Long, Shruti Bhilare, Vivek Kanhangad, Artur Costa-Pazo, Esteban Vázquez-Fernández, Daniel Pérez-Cabo, J. J. Moreira-Perez, Daniel González-Jiménez, A. Mohammadi, Sushil Bhattacharjee, Sébastien Marcel, S. Volkova, Y. Tang, N. Abe, L. Li, X. Feng, Z. Xia, X. Jiang, S. Liu, R. Shao, Pong C. Yuen, Waldir R. de Almeida, Fernanda A. Andaló, Rafael Padilha, Gabriel Bertocco, William Dias, Jacques Wainer, Ricardo da Silva Torres, Anderson Rocha, Marcus A. Angeloni, Guilherme Folego, Alan Godoy, and Abdenour Hadid. A competition on generalized software-based face presentation attack detection in mobile scenarios. In *IJCB*, pages 688–696. IEEE, 2017.

[14] Zinelabdine Boulkenafet, Jukka Komulainen, Lei Li, Xiaoyi Feng, and Abdenour Hadid. OULU-NPU: A mobile face presentation attack database with real-world variations. In *FG*, pages 612–618. IEEE Computer Society, 2017.

[15] Zinelabdine Boulkenafet, Jukka Komulainen, Lei Li, Xiaoyi Feng, and Abdenour Hadid. OULU-NPU: A mobile face presentation attack database with real-world variations. In *12th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2017, Washington, DC, USA, May 30 - June 3, 2017*, pages 612–618. IEEE Computer Society, 2017.

[16] Zinelabidine Boulkenafet, Jukka Komulainen, and Abdenour Hadid. Face anti-spoofing based on color texture analysis. In *2015 IEEE International Conference on Image Processing, ICIP 2015, Quebec City, QC, Canada, September 27-30, 2015*, pages 2636–2640. IEEE, 2015.

[17] Zinelabidine Boulkenafet, Jukka Komulainen, and Abdenour Hadid. Face spoofing detection using colour texture analysis. *IEEE Trans. Inf. Forensics Secur.*, 11(8):1818–1830, 2016.

[18] Fadi Boutros, Naser Damer, Meiling Fang, Florian Kirchbuchner, and Arjan Kuijper. Mixfacenets: Extremely efficient face recognition networks. In *International IEEE Joint Conference on Biometrics, IJCB 2021, Shenzhen, China, August 4-7, 2021*, pages 1–8. IEEE, 2021.

[19] Fadi Boutros, Naser Damer, Meiling Fang, Kiran B. Raja, Florian Kirchbuchner, and Arjan Kuijper. Compact models for periocular verification through knowledge distillation. In *BIOSIG 2020 - Proceedings of the 19th International Conference of the Biometrics Special Interest Group, online, 16.-18. September 2020*, volume P-306 of *LNI*, pages 291–298. Gesellschaft für Informatik e.V., 2020.

[20] Fadi Boutros, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Eye-mms: Miniature multi-scale segmentation network of key eye-regions in embedded applications. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 10 2019.

[21] Fadi Boutros, Meiling Fang, Marcel Klemt, Biying Fu, and Naser Damer. CR-FIQA: face image quality assessment by learning sample relative classifiability. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, Canada, June 18-22, 2023*, pages 1–10. IEEE (to appear), 2023.

[22] Fadi Boutros, Marco Huber, Patrick Siebke, Tim Rieber, and Naser Damer. Sface: Privacy-friendly and accurate face recognition using synthetic data. In *IJCB*, pages 1–11. IEEE, 2022.

[23] Fadi Boutros, Olga Kaehm, Meiling Fang, Florian Kirchbuchner, Naser Damer, and Arjan Kuijper. Low-resolution iris recognition via knowledge transfer. In *2022 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–5. IEEE, 2022.

[24] Fadi Boutros, Marcel Klemt, Meiling Fang, Arjan Kuijper, and Naser Damer. Unsupervised face recognition using unlabeled synthetic data. In *FG*, pages 1–8. IEEE, 2023.

[25] Martins Bruveris, Pouria Mortazavian, Jochem Gietema, and Mohan Mahadevan. Reducing geographic performance differentials for face recognition. In *WACV Workshops*, pages 98–106. IEEE, 2020.

[26] Han Cai, Ligeng Zhu, and Song Han. Proxylessnas: Direct neural architecture search on target task and hardware. In *ICLR (Poster)*. OpenReview.net, 2019.

[27] Rizhao Cai, Haoliang Li, Shiqi Wang, Changsheng Chen, and Alex C. Kot. DRL-FAS: A novel framework based on deep reinforcement learning for face anti-spoofing. *IEEE Trans. Inf. Forensics Secur.*, 16:937–951, 2021.

[28] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *13th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2018, Xi'an, China, May 15-19, 2018*, pages 67–74. IEEE Computer Society, 2018.

[29] Jacqueline G. Cavazos, P. Jonathon Phillips, Carlos Domingo Castillo, and Alice J. O'Toole. Accuracy comparison across face recognition algorithms: Where are we on measuring race bias? *IEEE Trans. Biom. Behav. Identity Sci.*, 3(1):101–111, 2021.

[30] Christophe Champod and Massimo Tistarelli. Biometric technologies for forensic science and policing: State of the art. In Massimo Tistarelli and Christophe Champod, editors, *Handbook of Biometrics for Forensic Science*, Advances in Computer Vision and Pattern Recognition, pages 1–15. Springer, 2017.

[31] Chaos Computer Club. Chaos computer club breaks apple touchid, 2013. `https://www.ccc.de/en/updates/2013/ccc-breaks-apple-touchid`.

[32] Prosenjit Chatterjee, Alex Yalchin, Joseph Shelton, Kaushik Roy, Xiaohong Yuan, and Kossi D. Edoh. Presentation attack detection using wavelet transform and deep residual neural net. In Guojun Wang, Jun Feng, Md. Zakirul Alam Bhuiyan, and Rongxing Lu, editors, *Security, Privacy, and Anonymity in Computation, Communication, and Storage - SpaCCS 2019 International Workshops, Atlanta, GA, USA, July 14-17, 2019, Proceedings*, volume 11637 of *Lecture Notes in Computer Science*, pages 86–94. Springer, 2019.

[33] Baoliang Chen, Wenhan Yang, Haoliang Li, Shiqi Wang, and Sam Kwong. Camera invariant feature learning for generalized face anti-spoofing. *IEEE Trans. Inf. Forensics Secur.*, 16:2477–2492, 2021.

[34] Baoliang Chen, Wenhan Yang, Haoliang Li, Shiqi Wang, and Sam Kwong. Camera invariant feature learning for generalized face anti-spoofing. *IEEE Trans. Inf. Forensics Secur.*, 16:2477–2492, 2021.

[35] Baoliang Chen, Wenhan Yang, and Shiqi Wang. Face anti-spoofing by fusing high and low frequency features for advanced generalization capability. In *MIPR*, pages 199–204. IEEE, 2020.

[36] Cunjian Chen and Arun Ross. A multi-task convolutional neural network for joint iris detection and presentation attack detection. In *2018 IEEE Winter Applications of Computer Vision Workshops, WACV Workshops 2018, Lake Tahoe, NV, USA, March 15, 2018*, pages 44–51. IEEE Computer Society, 2018.

[37] Cunjian Chen and Arun Ross. An explainable attention-guided iris presentation attack detector. In *IEEE WACV Workshops 2021, Waikola, HI, USA, 2021*, pages 97–106. IEEE, 2021.

[38] Haonan Chen, Guosheng Hu, Zhen Lei, Yaowu Chen, Neil Martin Robertson, and Stan Z. Li. Attention-based two-stream convolutional networks for face spoofing detection. *IEEE Trans. Inf. Forensics Secur.*, 15:578–593, 2020.

[39] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. SCA-CNN: spatial and channel-wise attention in convolutional networks for image captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6298–6306. IEEE Computer Society, 2017.

[40] Sheng Chen, Yang Liu, Xiang Gao, and Zhen Han. Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. In *CCBR*, volume 10996 of *Lecture Notes in Computer Science*, pages 428–438. Springer, 2018.

[41] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E. Hinton. Big self-supervised models are strong semi-supervised learners. In *NeurIPS*, 2020.

[42] Yang Chen, Tian Wang, Jingjing Wang, Peng Shi, Guangcun Shan, and Hichem Snoussi. Towards good practices in face anti-spoofing: An image reconstruction based method. In *2019 Chinese Automation Congress (CAC)*, pages 4700–4705, 2019.

[43] Ivana Chingovska, André Anjos, and Sébastien Marcel. On the effectiveness of local binary patterns in face anti-spoofing. In *BIOSIG*, volume P-196 of *LNI*, pages 1–7. GI, 2012.

[44] Ivana Chingovska, André Anjos, and Sébastien Marcel. On the effectiveness of local binary patterns in face anti-spoofing. In Arslan Brömme and Christoph Busch, editors, *2012 BIOSIG - Proceedings of the International Conference of Biometrics Special Interest Group, Darmstadt, Germany, September 6-7, 2012*, volume P-196 of *LNI*, pages 1–7. GI, 2012.

[45] Ivana Chingovska, Nesli Erdogmus, André Anjos, and Sébastien Marcel. Face recognition systems under spoofing attacks. In Thirimachos Bourlai, editor, *Face Recognition Across the Imaging Spectrum*, pages 165–194. Springer, 2016.

[46] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13 - 17, 2017*, pages 797–806. ACM, 2017.

[47] Elliot Creager, David Madras, Jörn-Henrik Jacobsen, Marissa A. Weis, Kevin Swersky, Toniann Pitassi, and Richard S. Zemel. Flexibly fair representation learning by disentanglement. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 1436–1445. PMLR, 2019.

[48] Adam Czajka and Kevin W. Bowyer. Presentation attack detection for iris recognition: An assessment of the state-of-the-art. *ACM Comput. Surv.*, 51(4):86:1–86:35, 2018.

[49] Adam Czajka, Zhaoyuan Fang, and Kevin W. Bowyer. Iris presentation attack detection based on photometric stereo features. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2019, Waikoloa Village, HI, USA, January 7-11, 2019*, pages 877–885. IEEE, 2019.

[50] Naser Damer, Fadi Boutros, Marius Süßmilch, Meiling Fang, Florian Kirchbuchner, and Arjan Kuijper. Masked face recognition: Human versus machine. *IET Biom.*, 7(1):15–26, 2022.

[51] Naser Damer, Fadi Boutros, Marius Süßmilch, Florian Kirchbuchner, and Arjan Kuijper. Extended evaluation of the effect of real and simulated masks on face recognition performance. *IET Biometrics*, 10(5):548–561, 2021.

[52] Naser Damer and Kristiyan Dimitrov. Practical view on face presentation attack detection. In Richard C. Wilson, Edwin R. Hancock, and William A. P. Smith,

editors, *Proceedings of the British Machine Vision Conference 2016, BMVC 2016, York, UK, September 19-22, 2016*. BMVA Press, 2016.

[53] Naser Damer, Meiling Fang, Patrick Siebke, Jan Niklas Kolf, Marco Hube, and Fadi Boutros. Mordiff: Recognition vulnerability and attack detectability of face morphing attacks created by diffusion autoencoders. In *11th IEEE International Workshop on Biometrics and Forensics, IWBF 2023, Barcelona, Spain, April 19-20, 2023*, pages 1–6. IEEE (to appear), 2023.

[54] Naser Damer, Jonas Henry Grebe, Cong Chen, Fadi Boutros, Florian Kirchbuchner, and Arjan Kuijper. The effect of wearing a mask on face recognition performance: an exploratory study. In *BIOSIG*, volume P-306, pages 1–10. Gesellschaft für Informatik e.V., 2020.

[55] Naser Damer, César Augusto Fontanillo López, Meiling Fang, Noémie Spiller, Minh Vu Pham, and Fadi Boutros. Privacy-friendly synthetic data for the development of face morphing attack detectors. In *CVPR Workshops*, pages 1605–1616. IEEE, 2022.

[56] Naser Damer, César Augusto Fontanillo López, Meiling Fang, Noémie Spiller, Minh Vu Pham, and Fadi Boutros. Privacy-friendly Synthetic Data for the Development of Face Morphing Attack Detectors. In *CVPR Workshops*, pages 1605–1616. IEEE, 2022.

[57] Naser Damer, Alexander Opel, and Alexander Nouak. Biometric source weighting in multi-biometric fusion: Towards a generalized and robust solution. In *22nd European Signal Processing Conference, EUSIPCO 2014, Lisbon, Portugal, September 1-5, 2014*, pages 1382–1386. IEEE, 2014.

[58] Naser Damer, Kiran B. Raja, Marius Süßmilch, Sushma Venkatesh, Fadi Boutros, Meiling Fang, Florian Kirchbuchner, Raghavendra Ramachandra, and Arjan Kuijper. ReGenMorph: Visibly Realistic GAN Generated Face Morphing Attacks by Attack Re-generation. In *ISVC (1)*, volume 13017 of *Lecture Notes in Computer Science*, pages 251–264. Springer, 2021.

[59] Naser Damer, Noémie Spiller, Meiling Fang, Fadi Boutros, Florian Kirchbuchner, and Arjan Kuijper. PW-MAD: Pixel-Wise Supervision for Generalized Face Morphing Attack Detection. In *ISVC (1)*, volume 13017 of *Lecture Notes in Computer Science*, pages 291–304. Springer, 2021.

[60] David Danks and Alex John London. Algorithmic bias in autonomous systems. In Carles Sierra, editor, *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 4691–4697. ijcai.org, 2017.

[61] Priyanka Das, Joseph McGrath, Zhaoyuan Fang, Aidan Boyd, Ganghee Jang, Amir Mohammadi, Sandip Purnapatra, David Yambay, Sébastien Marcel, Mateusz Trokielewicz, Piotr Maciejewicz, Kevin W. Bowyer, Adam Czajka, Stephanie Schuckers, Juan E. Tapia, Sebastian Gonzalez, Meiling Fang, Naser Damer, Fadi Boutros, Arjan Kuijper, Renu Sharma, Cunjian Chen, and Arun Ross. Iris liveness detection competition (livdet-iris) - the 2020 edition. In *IEEE IJCB, online, 28.09-01.10, 2020*, pages 1–9. IEEE, 2020.

[62] John Daugman. How iris recognition works. *IEEE Trans. Circuits Syst. Video Techn.*, 14(1):21–30, 2004.

[63] Tiago de Freitas Pereira, André Anjos, José Mario De Martino, and Sébastien Marcel. LBP - TOP based countermeasure against face spoofing attacks. In Jong-Il Park and Junmo Kim, editors, *Computer Vision - ACCV 2012 Workshops - ACCV 2012 International Workshops, Daejeon, Korea, November 5-6, 2012, Revised Selected Papers, Part I*, volume 7728 of *Lecture Notes in Computer Science*, pages 121–132. Springer, 2012.

[64] Tiago de Freitas Pereira and Sébastien Marcel. Fairness in biometrics: A figure of merit to assess biometric verification systems. *IEEE Trans. Biom. Behav. Identity Sci.*, 4(1):19–29, 2022.

[65] Debayan Deb and Anil K. Jain. Look locally infer globally: A generalizable face anti-spoofing approach. *IEEE Trans. Inf. Forensics Secur.*, 16:1143–1157, 2021.

[66] Jose Sanchez del Rio Saez, Daniela Moctezuma, Cristina Conde, Isaac Martín de Diego, and Enrique Cabello. Automated border control e-gates and facial recognition systems. *Comput. Secur.*, 62:49–72, 2016.

[67] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. IEEE Computer Society, 2009.

[68] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *IEEE Conference on Computer*

Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, pages 4690–4699. Computer Vision Foundation / IEEE, 2019.

[69] Pawel Drozdowski, Christian Rathgeb, Antitza Dantcheva, Naser Damer, and Christoph Busch. Demographic bias in biometrics: A survey on an emerging challenge. *CoRR*, abs/2003.02488, 2020.

[70] Gizem Erdogan and Arun Ross. Automatic detection of non-cosmetic soft contact lenses in ocular images. In Ioannis Kakadiaris, Walter J. Scheirer, and Laurence G. Hassebrook, editors, *Biometric and Surveillance Technology for Human and Activity Identification X*, volume 8712, pages 62 – 76. International Society for Optics and Photonics, SPIE, 2013.

[71] Nesli Erdogmus and Sébastien Marcel. Spoofing in 2d face recognition with 3d masks and anti-spoofing with kinect. In *IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems, BTAS 2013, Arlington, VA, USA, September 29 - October 2, 2013*, pages 1–6. IEEE, 2013.

[72] Meiling Fang, Hamza Ali, Arjan Kuijper, and Naser Damer. PatchSwap: Boosting the Generalizability of Face Presentation Attack Detection by Identity-aware Patch Swapping. In *International IEEE Joint Conference on Biometrics, IJCB 2022, Abu Dhabi, United Arab Emirates, October 10-13, 2022*, pages 1–8. IEEE, 2022.

[73] Meiling Fang, Fadi Boutros, and Naser Damer. Unsupervised face morphing attack detection via self-paced anomaly detection. In *International IEEE Joint Conference on Biometrics, IJCB 2022, Abu Dhabi, United Arab Emirates, October 10-13, 2022*, pages 1–8. IEEE, 2022.

[74] Meiling Fang, Fadi Boutros, and Naser Damer. Intra and cross-spectrum iris presentation attack detection in the nir and visible domains. In Sébastien Marcel, Julian Fierrez, and Nicholas Evans, editors, *Handbook of Biometric Anti-Spoofing: Presentation Attack Detection and Vulnerability Assessment*, pages 171–199, Singapore, 2023. Springer Nature Singapore.

[75] Meiling Fang, Fadi Boutros, Arjan Kuijper, and Naser Damer. Partial attack supervision and regional weighted inference for masked face presentation attack detection. In *16th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2021, Jodhpur, India, December 15-18, 2021*, pages 1–8. IEEE, 2021.

[76] Meiling Fang and Naser Damer. Face presentation attack detection by excavating causal clues and adapting embedding statistics. *CoRR*, abs/, 2023.

[77] Meiling Fang, Naser Damer, Fadi Boutros, Florian Kirchbuchner, and Arjan Kuijper. Deep learning multi-layer fusion for an accurate iris presentation attack detection. In *IEEE 23rd FUSION, online, July 6-9, 2020*, pages 1–8. IEEE, 2020.

[78] Meiling Fang, Naser Damer, Fadi Boutros, Florian Kirchbuchner, and Arjan Kuijper. Cross-database and cross-attack iris presentation attack detection using micro stripes analyses. *Image Vis. Comput.*, 105:104057, 2021.

[79] Meiling Fang, Naser Damer, Fadi Boutros, Florian Kirchbuchner, and Arjan Kuijper. Iris presentation attack detection by attention-based and deep pixel-wise binary supervision network. In *International IEEE Joint Conference on Biometrics, IJCB 2021, Shenzhen, China, August 4-7, 2021*, pages 1–8. IEEE, 2021.

[80] Meiling Fang, Naser Damer, Fadi Boutros, Florian Kirchbuchner, and Arjan Kuijper. The overlapping effect and fusion protocols of data augmentation techniques in iris PAD. *Mach. Vis. Appl.*, 33(1):8, 2022.

[81] Meiling Fang, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Demographic bias in presentation attack detection of iris recognition systems. In *28th European Signal Processing Conference, EUSIPCO 2020, Amsterdam, Netherlands, January 18-21, 2021*, pages 835–839. IEEE, 2020.

[82] Meiling Fang, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Micro stripes analyses for iris presentation attack detection. In *2020 IEEE International Joint Conference on Biometrics, IJCB 2020, online, Sep.28 - Oct.1, 2020*, pages 1–10. IEEE, 2020.

[83] Meiling Fang, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Learnable multi-level frequency decomposition and hierarchical attention mechanism for generalized face presentation attack detection. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, Waikoloa, HI, USA, January 3-8, 2022*, pages 1131–1140. IEEE, 2022.

[84] Meiling Fang, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Real masks and spoof faces: On the masked face presentation attack detection. *Pattern Recognition*, 123:108398, 2022.

[85] Meiling Fang, Marco Huber, and Naser Damer. SynthASpoof: Developing Face Presentation Attack Detection Based on Privacy-friendly Synthetic Data. In *CVPR Workshops*, pages 1605–1616. IEEE, 2023.

[86] Meiling Fang, Wufei Yang, Arjan Kuijper, Vitomir Struc, and Naser Damer. Fairness in face presentation attack detection. *CoRR*, abs/2112.06592, 2022.

[87] Soroush Fatemifar, Shervin Rahimzadeh Arashloo, Muhammad Awais, and Josef Kittler. Client-specific anomaly detection for face presentation attack detection. *Pattern Recognition*, 112:107696, 2021.

[88] Haocheng Feng, Zhibin Hong, Haixiao Yue, Yang Chen, Keyao Wang, Junyu Han, Jingtuo Liu, and Errui Ding. Learning generalized spoof cues for face anti-spoofing. *CoRR*, abs/2005.03922, 2020.

[89] Pedro M. Ferreira, Ana F. Sequeira, Diogo Pernes, Ana Rebelo, and Jaime S. Cardoso. Adversarial learning for a robust iris presentation attack detection method against unseen attack presentations. In Arslan Brömme, Christoph Busch, Antitza Dantcheva, Christian Rathgeb, and Andreas Uhl, editors, *2019 International Conference of the Biometrics Special Interest Group, BIOSIG 2019, Darmstadt, Germany, September 18-20, 2019*, pages 1–7. IEEE, 2019.

[90] Biying Fu, Cong Chen, Olaf Henniger, and Naser Damer. The relative contributions of facial parts qualities to the face image utility. *BIOSIG*, 2021.

[91] Biying Fu, Noemie Spiller, Cong Chen, and Naser Damer. The effect of face morphing on face image quality. *BIOSIG*, 2021.

[92] Javier Galbally, Sébastien Marcel, and Julian Fiérrez. Image quality assessment for fake biometric detection: Application to iris, fingerprint, and face recognition. *IEEE Trans. Image Process.*, 23(2):710–724, 2014.

[93] Anjith George and Sébastien Marcel. Deep pixel-wise binary supervision for face presentation attack detection. In *ICB*, pages 1–8. IEEE, 2019.

[94] Anjith George and Sébastien Marcel. Deep pixel-wise binary supervision for face presentation attack detection. In *2019 International Conference on Biometrics, ICB 2019, Crete, Greece, June 4-7, 2019*, pages 1–8. IEEE, 2019.

[95] Sixue Gong, Xiaoming Liu, and Anil K. Jain. Mitigating face recognition bias via group adaptive classifier. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 3414–3424. Computer Vision Foundation / IEEE, 2021.

[96] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2672–2680, 2014.

[97] Diego Gragnaniello, Carlo Sansone, and Luisa Verdoliva. Iris liveness detection for mobile devices based on local descriptors. *Pattern Recognition Letters*, 57.

[98] Patrick Grother. Demographic differentials in face recognition algorithms. Technical report, 2021.

[99] Patrick Grother, Mei Ngan, and Kayee Hanaoka. Face recognition vendor test (frvt) part 3: Demographic effects. Technical report, National Institute of Standards and Technology, 2019.

[100] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III*, volume 9907, pages 87–102. Springer, 2016.

[101] Mehak Gupta, Vishal Singh, Akshay Agarwal, Mayank Vatsa, and Richa Singh. Generalized iris presentation attack detection algorithm under cross-database settings. In *2020 IEEE ICPR, Milano Congress Center, Italy, 10-15, January 2021*. IEEE, 2020.

[102] Priyanshu Gupta, Shipra Behera, Mayank Vatsa, and Richa Singh. On iris spoofing using print attack. In *22nd International Conference on Pattern Recognition, ICPR 2014, Stockholm, Sweden, Aug. 24-28, 2014*, pages 1681–1686. IEEE Computer Society, 2014.

[103] Guy Hacohen and Daphna Weinshall. On the power of curriculum learning in training deep networks. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, pages 2535–2544. PMLR, 2019.

[104] George M. Haley and B. S. Manjunath. Rotation-invariant texture classification using a complete space-frequency model. *IEEE Trans. Image Process.*, 8(2):255–269, 1999.

[105] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *NIPS*, pages 3315–3323, 2016.

[106] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9726–9735. Computer Vision Foundation / IEEE, 2020.

[107] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778. IEEE Computer Society, 2016.

[108] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016.

[109] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, 1997.

[110] Steven Hoffman, Renu Sharma, and Arun Ross. Convolutional neural networks for iris presentation attack detection: Toward cross-dataset and cross-sensor generalization. In *2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 1620–1628. IEEE Computer Society, 2018.

[111] Steven Hoffman, Renu Sharma, and Arun Ross. Iris + ocular: Generalized iris presentation attack detection using multiple convolutional neural networks. In *2019 International Conference on Biometrics, ICB 2019, Crete, Greece, June 4-7, 2019*, pages 1–8. IEEE, 2019.

[112] Arne Holst. Smartphone users worldwide 2016-2021. Accessed: 2019-11-11.

[113] Qibin Hou, Daquan Zhou, and Jiashi Feng. Coordinate attention for efficient mobile network design. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 13713–13722. Computer Vision Foundation / IEEE, 2021.

[114] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, and Hartwig Adam. Searching for mobilenetv3. *CoRR*, abs/1905.02244, 2019.

[115] John J. Howard, Yevgeniy B. Sirotin, and Arun R. Vemury. The effect of broad and specific demographic homogeneity on the imposter distributions and false match rates in face recognition algorithm performance. In *10th IEEE International Conference on Biometrics Theory, Applications and Systems, BTAS 2019, Tampa, FL, USA, September 23-26, 2019*, pages 1–8. IEEE, 2019.

[116] Sheng-Hsun Hsieh, Yunghui Li, Wei Wang, and Chung-Hao Tien. A novel anti-spoofing solution for iris recognition toward cosmetic contact lens attack using spectral ica analysis. *Sensors*, 18:795, 03 2018.

[117] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 2261–2269. IEEE Computer Society, 2017.

[118] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[119] Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. Curricularface: Adaptive curriculum learning loss for deep face recognition. In *CVPR*, pages 5900–5909. Computer Vision Foundation / IEEE, 2020.

[120] Marco Huber, Fadi Boutros, Anh Thi Luu, Kiran B. Raja, Raghavendra Ramachandra, Naser Damer, Pedro C. Neto, Tiago Gonçalves, Ana F. Sequeira, Jaime S. Cardoso, João Tremoço, Miguel Lourenço, Sergio Serra, Eduardo Cermeño, Marija Ivanovska, Borut Batagelj, Andrej Kronovsek, Peter Peer, and Vitomir Struc. SYN-MAD 2022: Competition on face morphing attack detection based on privacy-aware synthetic training data. In *IJCB*, pages 1–10. IEEE, 2022.

[121] Ken Hughes and Kevin W. Bowyer. Detection of contact-lens-based iris biometric spoofs using stereo imaging. In *46th Hawaii International Conference on System Sciences, HICSS 2013, Wailea, HI, USA, January 7-10, 2013*, pages 1763–1772. IEEE Computer Society, 2013.

[122] Anas Husseis, Judith Liu-Jimenez, Ines Goicoechea-Telleria, and Raul Sánchez-Reillo. A survey in presentation attack and presentation attack detection. In *ICCST*, pages 1–13. IEEE, 2019.

[123] Ben Hutchinson and Margaret Mitchell. 50 years of test (un)fairness: Lessons for machine learning. In *FAT*, pages 49–58. ACM, 2019.

[124] Hiroshi Inoue. Data augmentation by pairing samples for images classification. *CoRR*, abs/1801.02929, 2018.

[125] International Civil Aviation Organization. Technical report: Portrait quality (reference facial images for mrtd), 2018.

[126] International Organization for Standardization. ISO/IEC DIS 30107-1:2016: Information Technology – Biometric presentation attack detection – P. 1: Framework, 2017.

[127] International Organization for Standardization. ISO/IEC DIS 30107-3:2016: Information Technology – Biometric presentation attack detection – P. 3: Testing and reporting, 2017.

[128] International Organization for Standardization. ISO/IEC TR 24027:2021 Information Technology – Artificial intelligence (AI) – Bias in AI systems and AI aided deicison making, 2021.

[129] International Organization for Standardization. ISO/IEC WD 19795-10: Biometric performance testing and reporting - Part 10: Quantifying biometric system performance variation across demographic groups, 2023.

[130] International Organization for Standardization and International Electrotechnical Committee. ISO/IEC 2382 37:2017: Information Technology - Vocabulary - Part 37: Biometrics, 2nd ed., February 2017.

[131] ISO/IEC JTC1 SC37 Biometrics. ISO/IEC 2382-37:2017 Information technology - Vocabulary - Part 37: Biometrics. International Organization for Standardization, 2017.

[132] ISO/IEC JTC1 SC37 Biometrics. ISO/IEC 19795-1:2021 Information technology — Biometric performance testing and reporting — Part 1: Principles and framework. International Organization for Standardization, 2021.

[133] Anil K. Jain, Arun Ross, and Salil Prabhakar. An introduction to biometric recognition. *IEEE Trans. Circuits Syst. Video Technol.*, 14(1):4–20, 2004.

[134] Shan Jia, Guodong Guo, and Zhengquan Xu. A survey on 3d mask presentation attack detection and countermeasures. *Pattern Recognition*, 98:107032, 2020.

[135] Shan Jia, Chuanbo Hu, Xin Li, and Zhengquan Xu. Face spoofing detection under super-realistic 3d wax face attacks. *Pattern Recognition Lett.*, 145:103–109, 2021.

[136] Shan Jia, Chuanbo Hu, Xin Li, and Zhengquan Xu. Face spoofing detection under super-realistic 3d wax face attacks. *Pattern Recognition Lett.*, 145:103–109, 2021.

[137] Shan Jia, Xin Li, Chuanbo Hu, Guodong Guo, and Zhengquan Xu. 3d face anti-spoofing with factorized bilinear coding. *IEEE Trans. Circuits Syst. Video Technol.*, 31(10):4031–4045, 2021.

[138] Shan Jia, Xin Li, Chuanbo Hu, and Zhengquan Xu. Spoofing and anti-spoofing with wax figure faces. *CoRR*, abs/1910.05457, 2019.

[139] Yunpei Jia, Jie Zhang, Shiguang Shan, and Xilin Chen. Single-side domain generalization for face anti-spoofing. In *CVPR*, pages 8481–8490. Computer Vision Foundation / IEEE, 2020.

[140] Yunpei Jia, Jie Zhang, Shiguang Shan, and Xilin Chen. Unified unsupervised and semi-supervised domain adaptation network for cross-scenario face anti-spoofing. *Pattern Recognition*, 115:107888, 2021.

[141] Amin Jourabloo, Yaojie Liu, and Xiaoming Liu. Face de-spoofing: Anti-spoofing via noise modeling. In *ECCV (13)*, volume 11217 of *Lecture Notes in Computer Science*, pages 297–315. Springer, 2018.

[142] James S. Doyle Jr. and Kevin W. Bowyer. Robust detection of textured contact lenses in iris recognition using BSIF. *IEEE Access*, 3:1672–1683, 2015.

[143] James S. Doyle Jr., Kevin W. Bowyer, and Patrick J. Flynn. Variation in accuracy of textured contact lens detection based on sensor and lens pattern. In *IEEE 6th BTAS, 2013, Arlington, VA, USA, Sep.29-Oct. 2, 2013*, pages 1–7. IEEE, 2013.

[144] Alperen Kantarci, Hasan Dertli, and Hazim Kemal Ekenel. Shuffled patch-wise supervision for presentation attack detection. *BIOSIG*, 2021.

[145] Ira Kemelmacher-Shlizerman, Steven M. Seitz, Daniel Miller, and Evan Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *CVPR*, pages 4873–4882. IEEE Computer Society, 2016.

[146] Portia Kentish. Is paying with your iris the future of transactions? polish start-up payeye certainly thinks so. Accessed: 2020-07-07.

[147] Taewook Kim, Yonghyun Kim, Inhan Kim, and Daijin Kim. BASN: enriching feature representation using bipartite auxiliary supervisions for face anti-spoofing. In *2019 IEEE/CVF International Conference on Computer Vision Workshops, ICCV Workshops 2019, Seoul, Korea (South), October 27-28, 2019*, pages 494–503. IEEE, 2019.

[148] Gabriela Y. Kimura, Diego Rafael Lucio, Alceu S. Britto Jr., and David Menotti. CNN hyperparameter tuning applied to iris liveness detection. *CoRR*, abs/2003.00833, 2020.

[149] Davis E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009.

[150] Brendan Klare, Mark James Burge, Joshua C. Klontz, Richard W. Vorder Bruegge, and Anil K. Jain. Face recognition performance: Role of demographic information. *IEEE Trans. Inf. Forensics Secur.*, 7(6):1789–1801, 2012.

[151] Naman Kohli, Daksha Yadav, Mayank Vatsa, and Richa Singh. Revisiting iris recognition with color cosmetic contact lenses. In *International Conference on Biometrics, ICB 2013, 4-7 June, 2013, Madrid, Spain*, pages 1–7. IEEE, 2013.

[152] Naman Kohli, Daksha Yadav, Mayank Vatsa, Richa Singh, and Afzel Noore. Detecting medley of iris spoofing attacks using DESIST. In *8th IEEE BTAS 2016, Niagara Falls, NY, USA, September 6-9, 2016*, pages 1–6. IEEE, 2016.

[153] György Kovács. An empirical comparison and evaluation of minority oversampling techniques on a large number of imbalanced datasets. *Appl. Soft Comput.*, 83, 2019.

[154] K. S. Krishnapriya, Vítor Albiero, Kushal Vangara, Michael C. King, and Kevin W. Bowyer. Issues related to face recognition accuracy varying based on race and skin tone. *IEEE Transactions on Technology and Society*, 1(1):8–20, 2020.

[155] A. Kuehlkamp, A. Pinto, A. Rocha, K. W. Bowyer, and A. Czajka. Ensemble of multi-view learning classifiers for cross-domain iris presentation attack detection. *IEEE TIFS*, 14(6):1419–1431, 6 2019.

[156] Mohit Kumar and Niladri B. Puhan. RANSAC lens boundary feature based kernel SVM for transparent contact lens detection. *IET Biom.*, 8(3):177–184, 2019.

[157] Eui Chul Lee, Kang Ryoung Park, and Jaihie Kim. Fake iris detection by using purkinje image. In David Zhang and Anil K. Jain, editors, *Advances in Biometrics,*

International Conference, ICB 2006, Hong Kong, China, January 5-7, 2006, Proceedings, volume 3832 of *Lecture Notes in Computer Science*, pages 397–403. Springer, 2006.

[158] S. J. Lee, K. R. Park, Y. J. Lee, K. Bae, and J. Kim. Multifeature-based fake iris detection method, 46(12):127204, 2007.

[159] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C. Kot. Domain generalization with adversarial feature learning. In *CVPR*, pages 5400–5409. IEEE Computer Society, 2018.

[160] Jiangwei Li, Yunhong Wang, Tieniu Tan, and Anil K. Jain. Live face detection based on the analysis for fourier spectra. In *Biometric Technology for Human Identification*, volume 5404, pages 296–303, 2004.

[161] Lei Li, Zhaoqiang Xia, Xiaoyue Jiang, Fabio Roli, and Xiaoyi Feng. Compactnet: learning a compact space for face presentation attack detection. *Neurocomputing*, 409:191–207, 2020.

[162] Sheng Li, Xun Zhu, Guorui Feng, Xinpeng Zhang, and Zhenxing Qian. Diffusing the liveness cues for face anti-spoofing. In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, pages 1636–1644. ACM, 2021.

[163] Xiang Li, Wenhai Wang, Xiaolin Hu, and Jian Yang. Selective kernel networks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 510–519. Computer Vision Foundation / IEEE, 2019.

[164] Lindsey O'Donnell. Researchers bypass apple faceid using biometrics 'achilles heel', 2019. `https://threatpost.com/researchers-bypass-apple-faceid-using-biometrics-achilles-heel/147109/`.

[165] Hanxiao Liu, Karen Simonyan, and Yiming Yang. DARTS: differentiable architecture search. In *ICLR (Poster)*. OpenReview.net, 2019.

[166] Siqi Liu, Xiangyuan Lan, and Pong C. Yuen. Remote photoplethysmography correspondence feature for 3d mask face presentation attack detection. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XVI*, volume 11220 of *Lecture Notes in Computer Science*, pages 577–594. Springer, 2018.

[167] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6738–6746. IEEE Computer Society, 2017.

[168] Yaojie Liu, Amin Jourabloo, and Xiaoming Liu. Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In *CVPR*, pages 389–398. IEEE Computer Society, 2018.

[169] Yaojie Liu, Joel Stehouwer, Amin Jourabloo, and Xiaoming Liu. Deep tree learning for zero-shot face anti-spoofing. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 4680–4689. Computer Vision Foundation / IEEE, 2019.

[170] Yaojie Liu, Joel Stehouwer, and Xiaoming Liu. On disentangling spoof trace for generic face anti-spoofing. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XVIII*, volume 12363 of *Lecture Notes in Computer Science*, pages 406–422. Springer, 2020.

[171] Ana Carolina Lorena and André Carlos Ponce de Leon Ferreira de Carvalho. Building binary-tree-based multiclass classifiers using separability measures. *Neurocomputing*, 73(16-18):2837–2845, 2010.

[172] Oeslle Lucena, Amadeu Junior, Vitor Moia, Roberto Souza, Eduardo Valle, and Roberto de Alencar Lotufo. Transfer learning using convolutional neural networks for face anti-spoofing. In Fakhri Karray, Aurélio Campilho, and Farida Cheriet, editors, *Image Analysis and Recognition - 14th International Conference, ICIAR 2017, Montreal, QC, Canada, July 5-7, 2017, Proceedings*, volume 10317, pages 27–34. Springer, 2017.

[173] Jukka Määttä, Abdenour Hadid, and Matti Pietikäinen. Face spoofing detection from single images using micro-texture analysis. In *IJCB*, pages 1–7. IEEE Computer Society, 2011.

[174] Hareesh Mandalapu, Raghavendra Ramachandra, and Christoph Busch. Empirical evaluation of texture-based print and contact lens iris presentation attack detection methods. In *Proceedings of the 3rd ICBEA 2019, Stockholm, Sweden, May 29-31, 2019*. ACM, 2019.

[175] Yoanna Martínez-Díaz, Luis S. Luevano, Heydi Mendez Vazquez, Miguel Nicolás-Díaz, Leonardo Chang, and Miguel González-Mendoza. Shufflefacenet: A lightweight face architecture for efficient and highly-accurate face recognition. In *ICCV Workshops*, pages 2721–2728. IEEE, 2019.

[176] Joseph McGrath, Kevin W. Bowyer, and Adam Czajka. Open source presentation attack detection baseline for iris recognition. *CoRR*, abs/1809.10172, 2018.

[177] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6):115:1–115:35, 2021.

[178] David Menotti, Giovani Chiachia, Allan da Silva Pinto, William Robson Schwartz, Hélio Pedrini, Alexandre Xavier Falcão, and Anderson Rocha. Deep representations for iris, face, and fingerprint spoofing detection. *IEEE Trans. Inf. Forensics Secur.*, 10(4):864–879, 2015.

[179] Amir Mohammadi, Sushil Bhattacharjee, and Sébastien Marcel. Deeply vulnerable: a study of the robustness of face recognition to presentation attacks. *IET Biom.*, 7(1):15–26, 2018.

[180] Amir Mohammadi, Sushil Bhattacharjee, and Sébastien Marcel. Improving cross-dataset performance of face presentation attack detection systems using face recognition datasets. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*, pages 2947–2951. IEEE, 2020.

[181] Mei L Ngan, Patrick J Grother, and Kayee K Hanaoka. Ongoing face recognition vendor test (frvt) part 6b: Face recognition accuracy with face masks using post-covid-19 algorithms. 2020.

[182] Dat Tien Nguyen, Tuyen Danh Pham, Young-Woo Lee, and Kang Ryoung Park. Deep learning-based enhanced presentation attack detection for iris recognition by combining features from local and global regions based on NIR camera sensor. *Sensors*, 18(8):2601, 2018.

[183] Timo Ojala, Matti Pietikäinen, and Topi Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(7):971–987, 2002.

[184] Tiago Freitas Pereira, Jukka Komulainen, André Anjos, José Mario De Martino, Abdenour Hadid, Matti Pietikäinen, and Sébastien Marcel. Face liveness detection using dynamic texture. *EURASIP J. Image Video Process.*, 2014:2, 2014.

[185] Sandip Purnapatra, Nic Smalt, Keivan Bahmani, Priyanka Das, David Yambay, Amir Mohammadi, Anjith George, Thirimachos Bourlai, Sébastien Marcel, Stephanie Schuckers, Meiling Fang, Naser Damer, Fadi Boutros, Arjan Kuijper, Alperen Kantarci, Basar Demir, Zafer Yildiz, Zabi Ghafoory, Hasan Dertli, Hazim Kemal Ekenel, Son Vu, Vassilis Christophides, Dashuang Liang, Guanghao Zhang, Zhanlong Hao, Junfu Liu, Yufeng Jin, Samo Liu, Samuel Huang, Salieri Kuei, Jag Mohan Singh, and Raghavendra Ramachandra. Face liveness detection competition (livdet-face) - 2021. In *IJCB*, pages 1–10. IEEE, 2021.

[186] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *ECCV (12)*, volume 12357 of *Lecture Notes in Computer Science*, pages 86–103. Springer, 2020.

[187] Haibo Qiu, Baosheng Yu, Dihong Gong, Zhifeng Li, Wei Liu, and Dacheng Tao. Synface: Face recognition with synthetic data. In *ICCV*, pages 10860–10870. IEEE, 2021.

[188] Ramachandra Raghavendra and Christoph Busch. Robust scheme for iris presentation attack detection using multiscale binarized statistical image features. *IEEE Trans. Information Forensics and Security*, 10(4):703–715, 2015.

[189] Ramachandra Raghavendra, Kiran B. Raja, and Christoph Busch. Contlensnet: Robust iris contact lens detection using deep convolutional neural networks. In *2017 IEEE Winter Conference on Applications of Computer Vision, WACV 2017, Santa Rosa, CA, USA, March 24-31, 2017*, pages 1160–1167. IEEE Computer Society, 2017.

[190] Kiran B. Raja, Ramachandra Raghavendra, and Christoph Busch. Presentation attack detection using laplacian decomposed frequency response for visible spectrum and near-infra-red iris systems. In *IEEE 7th International Conference on Biometrics Theory, Applications and Systems, BTAS 2015, Arlington, VA, USA, September 8-11, 2015*, pages 1–8. IEEE, 2015.

[191] Christian Rathgeb, Andreas Uhl, Peter Wild, and Heinz Hofbauer. Design decisions for an iris recognition sdk. In Kevin Bowyer and Mark J. Burge, editors, *Handbook of Iris Recognition*, Advances in Computer Vision and Pattern Recognition. Springer, second edition edition, 2016.

[192] Narsi Reddy, Ajita Rattani, and Reza Derakhshani. Comparison of deep learning models for biometric-based mobile user authentication. In *9th IEEE International Conference on Biometrics Theory, Applications and Systems, BTAS 2018, Redondo Beach, CA, USA, October 22-25, 2018*, pages 1–6. IEEE, 2018.

[193] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich, Germany, October 5 - 9, 2015, Proceedings, Part III*, volume 9351 of *Lecture Notes in Computer Science*, pages 234–241. Springer, 2015.

[194] Mohammad Rostami, Leonidas Spinoulas, Mohamed E. Hussein, Joe Mathai, and Wael Abd-Almageed. Detection and continual learning of novel face presentation attacks. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 14831–14840. IEEE, 2021.

[195] Mohammad Rostami, Leonidas Spinoulas, Mohamed E. Hussein, Joe Mathai, and Wael Abd-Almageed. Detection and continual learning of novel face presentation attacks. In *ICCV*, pages 14831–14840. IEEE, 2021.

[196] Ryne Hagher. Samsung galaxy s10 face unlock can be fooled by a photo, video, or even your sister, 2019. `https://www.androidpolice.com/2019/03/09/samsung-galaxy-s10-face-unlock-can-be-fooled-by-a-photo-video-or-even-your-sister/`.

[197] Krishnapriya K. S, Kushal Vangara, Michael C. King, Vitor Albiero, and Kevin W. Bowyer. Characterizing the variability in face recognition accuracy relative to race. In *CVPR Workshops*, pages 2278–2285. Computer Vision Foundation / IEEE, 2019.

[198] T Sabhanayagam, V Prasanna Venkatesan, and K Senthamaraikannan. A comprehensive survey on various biometric systems. *International Journal of Applied Engineering Research*, 13(5):2276–2297, 2018.

[199] José A. Sáez, Bartosz Krawczyk, and Michal Wozniak. Analyzing the oversampling of different classes and types of examples in multi-class imbalanced datasets. *Pattern Recognition*, 57:164–178, 2016.

[200] Suman Saha, Wenhao Xu, Menelaos Kanakis, Stamatios Georgoulis, Yuhua Chen, Danda Pani Paudel, and Luc Van Gool. Domain agnostic feature learning for image and video based face anti-spoofing. In *CVPR Workshops*, pages 3490–3499. IEEE, 2020.

[201] LTD. SAMSUNG ELECTRONICS CO. How does the iris scanner work on galaxy s9, galaxy s9+, and galaxy note9? Accessed: 2021-04-19.

[202] Reva Schwartz, Apostol Vassilev, Kristen Greene, Lori Perine, Adrew Burt, and Partick Hall. Towards a standard for identifying and managing bias in artificial itelligence. Technical report, 2022.

[203] Rui Shao, Xiangyuan Lan, Jiawei Li, and Pong C. Yuen. Multi-adversarial discriminative deep domain generalization for face presentation attack detection. In *CVPR*, pages 10023–10031. Computer Vision Foundation / IEEE, 2019.

[204] Rui Shao, Xiangyuan Lan, and Pong C. Yuen. Regularized fine-grained meta face anti-spoofing. In *AAAI*, pages 11974–11981. AAAI Press, 2020.

[205] Renu Sharma and Arun Ross. D-netpad: An explainable and interpretable iris presentation attack detector. In *2020 IEEE International Joint Conference on Biometrics, IJCB 2020, online, Sep.28-Oct.1, 2020*, pages 1–10. IEEE, 2020.

[206] Pedro Silva, Eduardo José da S. Luz, Rafael Baeta, Hélio Pedrini, Alexandre Xavier Falcão, and David Menotti. An approach to iris contact lens detection based on deep image representations. In *28th SIBGRAPI Conference on Graphics, Patterns and Images, SIBGRAPI 2015, Salvador, Bahia, Brazil, August 26-29, 2015*, pages 157–164. IEEE Computer Society, 2015.

[207] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[208] Richa Singh, Puspita Majumdar, Surbhi Mittal, and Mayank Vatsa. Anatomizing bias in facial analysis. In *AAAI*, pages 12351–12358. AAAI Press, 2022.

[209] Philip Smith and Karl Ricanek. Mitigating algorithmic bias: Evolving an augmentation policy that is non-biasing. In *WACV Workshops*, pages 90–97. IEEE, 2020.

[210] Xiao Song, Xu Zhao, Liangji Fang, and Tianwei Lin. Discriminative representation combinations for accurate face spoofing detection. *Pattern Recognition*, 85:220–231, 2019.

[211] Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. Curriculum self-paced learning for cross-domain object detection. *Comput. Vis. Image Underst.*, 204:103166, 2021.

[212] José Augusto Stuchi, Marcus A. Angeloni, Rodrigo F. Pereira, Levy Boccato, Guilherme Folego, Paulo Victor de Souza Prado, and Romis Ribeiro Faissol Attux. Improving image classification with frequency domain layers for feature extraction. In *MLSP*, pages 1–6. IEEE, 2017.

[213] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2818–2826. IEEE Computer Society, 2016.

[214] Xiaoyang Tan, Yi Li, Jun Liu, and Lin Jiang. Face liveness detection from a single image with sparse low rank bilinear discriminative model. In Kostas Daniilidis, Petros Maragos, and Nikos Paragios, editors, *Computer Vision - ECCV 2010 - 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part VI*, volume 6316, pages 504–517. Springer, 2010.

[215] Luke Taylor and Geoff Nitschke. Improving deep learning using generic data augmentation. *CoRR*, abs/1708.06020, 2017.

[216] Philipp Terhörst, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Face quality estimation and its correlation to demographic and non-demographic bias in face recognition. In *IJCB*, pages 1–11. IEEE, 2020.

[217] Philipp Terhörst, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. SER-FIQ: unsupervised estimation of face image quality based on stochastic embedding robustness. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 5650–5659, 2020.

[218] Philipp Terhörst, Jan Niklas Kolf, Marco Huber, Florian Kirchbuchner, Naser Damer, Aythami Morales, Julian Fiérrez, and Arjan Kuijper. A comprehensive study on face recognition biases beyond demographics. *CoRR*, abs/2103.01592, 2021.

[219] Philipp Terhörst, Mai Ly Tran, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Comparison-level mitigation of ethnic bias in face recognition. In *IWBF*, pages 1–6. IEEE, 2020.

[220] Tom Nardi. Face id defeated with 3d printed mask, 2017. https://hackaday.com/2017/11/14/face-id-defeated-with-3d-printed-mask-maybe/.

[221] Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. In *CVPR*, pages 1521–1528. IEEE Computer Society, 2011.

[222] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

[223] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017.

[224] Sahil Verma and Julia Rubin. Fairness definitions explained. In *FairWare@ICSE*, pages 1–7. ACM, 2018.

[225] Raul Vicente-Garcia, Lukasz Wandzik, Louisa Grabner, and Jörg Krüger. The harms of demographic bias in deep face recognition research. In *2019 International Conference on Biometrics, ICB 2019, Crete, Greece, June 4-7, 2019*, pages 1–6. IEEE, 2019.

[226] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2020, Seattle, WA, USA, June 14-19, 2020*, pages 111–119. Computer Vision Foundation / IEEE, 2020.

[227] Mei Wang and Weihong Deng. Mitigating bias in face recognition using skewness-aware reinforcement learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9322–9331, 2020.

[228] Mei Wang and Weihong Deng. Mitigating bias in face recognition using skewness-aware reinforcement learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 9319–9328. Computer Vision Foundation / IEEE, 2020.

[229] Xin Wang, Yudong Chen, and Wenwu Zhu. A survey on curriculum learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(9):4555–4576, 2022.

[230] Yu-Chun Wang, Chien-Yi Wang, and Shang-Hong Lai. Disentangled representation with dual-stage feature learning for face anti-spoofing. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, Waikoloa, HI, USA, January 3-8, 2022*, pages 1234–1243. IEEE, 2022.

[231] Zezheng Wang, Chenxu Zhao, Yunxiao Qin, Qiusheng Zhou, and Zhen Lei. Exploiting temporal and depth information for multi-frame face anti-spoofing. *CoRR*, abs/1811.05118, 2018.

[232] Zhuoshi Wei, Xianchao Qiu, Zhenan Sun, and Tieniu Tan. Counterfeit iris detection based on texture analysis. In *19th ICPR, Dec. 8-11, 2008, Florida, USA*, pages 1–4. IEEE Computer Society, 2008.

[233] Di Wen, Hu Han, and Anil K. Jain. Face spoof detection with image distortion analysis. *IEEE Trans. Inf. Forensics Secur.*, 10(4):746–761, 2015.

[234] Di Wen, Hu Han, and Anil K. Jain. Face spoof detection with image distortion analysis. *IEEE Trans. Inf. Forensics Secur.*, 10(4):746–761, 2015.

[235] Lior Wolf, Tal Hassner, and Itay Maoz. Face recognition in unconstrained videos with matched background similarity. In *The 24th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, Colorado Springs, CO, USA, 20-25 June 2011*, pages 529–534. IEEE Computer Society, 2011.

[236] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: convolutional block attention module. In *Computer Vision - 15 th ECCV 2018, Munich, Germany, September 8-14, 2018, Proceedings, Part VII*, volume 11211 of *Lecture Notes in Computer Science*, pages 3–19. Springer, 2018.

[237] Zheng Xu, Wen Li, Li Niu, and Dong Xu. Exploiting low-rank structure from latent domains for domain generalization. In *ECCV (3)*, volume 8691 of *Lecture Notes in Computer Science*, pages 628–643. Springer, 2014.

[238] Daksha Yadav, Naman Kohli, Akshay Agarwal, Mayank Vatsa, Richa Singh, and Afzel Noore. Fusion of handcrafted and deep learning features for large-scale multiple iris presentation attack detection. In *2018 IEEE CVPR Workshops, Salt Lake City, USA, June 18-22, 2018*, pages 572–579. IEEE Computer Society, 2018.

[239] Daksha Yadav, Naman Kohli, James S. Doyle Jr., Richa Singh, Mayank Vatsa, and Kevin W. Bowyer. Unraveling the effect of textured contact lenses on iris recognition. *IEEE TIFS*, 9(5):851–862, 2014.

[240] Daksha Yadav, Naman Kohli, Mayank Vatsa, Richa Singh, and Afzel Noore. Unconstrained visible spectrum iris with textured contact lens variations: Database and benchmarking. In *2017 IEEE International Joint Conference on Biometrics, IJCB 2017, Denver, CO, USA, October 1-4, 2017*, pages 574–580. IEEE, 2017.

[241] Daksha Yadav, Naman Kohli, Mayank Vatsa, Richa Singh, and Afzel Noore. Iris presentation attack via textured contact lens in unconstrained environment. In *2018 IEEE Winter Conference on Applications of Computer Vision, WACV 2018, Lake Tahoe, NV, USA, March 12-15, 2018*, pages 503–511. IEEE Computer Society, 2018.

[242] Daksha Yadav, Naman Kohli, Mayank Vatsa, Richa Singh, and Afzel Noore. Detecting textured contact lens in uncontrolled environment using densepad. In *IEEE CVPR Workshops 2019, Long Beach, USA, June 16-20, 2019*, page 1. IEEE, 2019.

[243] David Yambay, Benedict Becker, Naman Kohli, Daksha Yadav, Adam Czajka, Kevin W. Bowyer, Stephanie Schuckers, Richa Singh, Mayank Vatsa, Afzel Noore, Diego Gragnaniello, Carlo Sansone, Luisa Verdoliva, Lingxiao He, Yiwei Ru, Haiqing Li, Nianfeng Liu, Zhenan Sun, and Tieniu Tan. Livdet iris 2017 - iris liveness detection competition 2017. In *IEEE IJCB, Denver, CO, USA, October 1-4, 2017*, pages 733–741. IEEE, 2017.

[244] David Yambay, Brian Walczak, Stephanie Schuckers, and Adam Czajka. Livdet-iris 2015 - iris liveness detection competition 2015. In *IEEE International Conference on Identity, Security and Behavior Analysis, ISBA 2017, New Delhi, India, February 22-24, 2017*, pages 1–6. IEEE, 2017.

[245] Xiao Yang, Wenhan Luo, Linchao Bao, Yuan Gao, Dihong Gong, Shibao Zheng, Zhifeng Li, and Wei Liu. Face anti-spoofing: Model matters, so does data. In *CVPR*, pages 3507–3516. Computer Vision Foundation / IEEE, 2019.

[246] Zitong Yu, Xiaobai Li, Xuesong Niu, Jingang Shi, and Guoying Zhao. Face anti-spoofing with human material perception. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part VII*, volume 12352 of *Lecture Notes in Computer Science*, pages 557–575. Springer, 2020.

[247] Zitong Yu, Xiaobai Li, Jingang Shi, Zhaoqiang Xia, and Guoying Zhao. Revisiting pixel-wise supervision for face anti-spoofing. *IEEE Trans. Biom. Behav. Identity Sci.*, 3(3):285–295, 2021.

[248] Zitong Yu, Yunxiao Qin, Xiaobai Li, Chenxu Zhao, Zhen Lei, and Guoying Zhao. Deep learning for face anti-spoofing: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(1):1–8, 2022.

[249] Zitong Yu, Yunxiao Qin, Hengshuang Zhao, Xiaobai Li, and Guoying Zhao. Dual-cross central difference network for face anti-spoofing. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 1281–1287. ijcai.org, 2021.

[250] Zitong Yu, Jun Wan, Yunxiao Qin, Xiaobai Li, Stan Z. Li, and Guoying Zhao. NAS-FAS: static-dynamic central difference network search for face anti-spoofing. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(9):3005–3023, 2021.

[251] Zitong Yu, Chenxu Zhao, Zezheng Wang, Yunxiao Qin, Zhuo Su, Xiaobai Li, Feng Zhou, and Guoying Zhao. Searching central difference convolutional networks for face anti-spoofing. In *CVPR*, pages 5294–5304. IEEE, 2020.

[252] Richard S. Zemel, Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. Learning fair representations. In *ICML (3)*, volume 28 of *JMLR Workshop and Conference Proceedings*, pages 325–333. JMLR.org, 2013.

[253] Han Zhang, Ian J. Goodfellow, Dimitris N. Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 7354–7363. PMLR, 2019.

[254] Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[255] Hui Zhang, Zhenan Sun, and Tieniu Tan. Contact lens detection based on weighted LBP. In *20th ICPR 2010, Istanbul, Turkey, August 2010*, pages 4279–4282. IEEE Computer Society, 2010.

[256] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Process. Lett.*, 23(10):1499–1503, 2016.

[257] Ke-Yue Zhang, Taiping Yao, Jian Zhang, Ying Tai, Shouhong Ding, Jilin Li, Feiyue Huang, Haichuan Song, and Lizhuang Ma. Face anti-spoofing via disentangled representation learning. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XIX*, volume 12364 of *Lecture Notes in Computer Science*, pages 641–657. Springer, 2020.

[258] Shifeng Zhang, Ajian Liu, Jun Wan, Yanyan Liang, Guodong Guo, Sergio Escalera, Hugo Jair Escalante, and Stan Z. Li. CASIA-SURF: A large-scale multi-modal benchmark for face anti-spoofing. *IEEE Trans. Biom. Behav. Identity Sci.*, 2(2):182–193, 2020.

[259] Yuanhan Zhang, Zhenfei Yin, Yidong Li, Guojun Yin, Junjie Yan, Jing Shao, and Ziwei Liu. Celeba-spoof: Large-scale face anti-spoofing dataset with rich annotations. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XII*, volume 12357, pages 70–85. Springer, 2020.

[260] Yuanhan Zhang, Zhenfei Yin, Yidong Li, Guojun Yin, Junjie Yan, Jing Shao, and Ziwei Liu. Celeba-spoof: Large-scale face anti-spoofing dataset with rich annotations. In *ECCV (12)*, volume 12357 of *Lecture Notes in Computer Science*, pages 70–85. Springer, 2020.

[261] Zhiwei Zhang, Junjie Yan, Sifei Liu, Zhen Lei, Dong Yi, and Stan Z. Li. A face antispoofing database with diverse attacks. In *ICB*, pages 26–31. IEEE, 2012.

[262] Zhiwei Zhang, Junjie Yan, Sifei Liu, Zhen Lei, Dong Yi, and Stan Z. Li. A face anti-spoofing database with diverse attacks. In Anil K. Jain, Arun Ross, Salil Prabhakar, and Jaihie Kim, editors, *5th IAPR International Conference on Biometrics, ICB 2012, New Delhi, India, March 29 - April 1, 2012*, pages 26–31. IEEE, 2012.

[263] Wei Zheng, Mengyuan Yue, Shuhuan Zhao, and Shuaiqi Liu. Attention-based spatial-temporal multi-scale network for face anti-spoofing. *IEEE Trans. Biom. Behav. Identity Sci.*, 3(3):296–307, 2021.

[264] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets V2: more deformable, better results. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 9308–9316. Computer Vision Foundation / IEEE, 2019.