Unmasking the imposters: towards improving the generalisation of deep learning methods for face presentation attack detection.

ABDULLAKUTTY, F.C.

2023

Unmasking the Imposters: Towards improving the generalisation of deep learning methods for face presentation attack detection

Faseela Chakkalakkal Abdullakutty

A report submitted as part of the requirements for the degree
of Doctor of Philosophy
at the School of Computing
Robert Gordon University
Aberdeen, Scotland

October 2023

Supervisor Prof. Eyad Elyan

Abstract

Identity theft has had a detrimental impact on the reliability of face recognition, which has been extensively employed in security applications. The most prevalent are presentation attacks. By using a photo, video, or mask of an authorized user, attackers can bypass face recognition systems. Fake presentation attacks are detected by the camera sensors of face recognition systems using face presentation attack detection. Presentation attacks can be detected using convolutional neural networks, commonly used in computer vision applications.

An in-depth analysis of current deep learning methods is used in this research to examine various aspects of detecting face presentation attacks. A number of new techniques are implemented and evaluated in this study, including pre-trained models, manual feature extraction, and data aggregation. The thesis explores the effectiveness of various machine learning and deep learning models in improving detection performance by using publicly available datasets with different dataset partitions than those specified in the official dataset protocol. Furthermore, the research investigates how deep models and data aggregation can be used to detect face presentation attacks, as well as a novel approach that combines manual features with deep features in order to improve detection accuracy. Moreover, task-specific features are also extracted using pre-trained deep models to enhance the performance of detection and generalisation further.

This problem is motivated by the need to achieve generalization against new and rapidly evolving attack variants. It is possible to extract identifiable features from presentation attack variants in order to detect them. However, new methods are needed to deal with emerging attacks and improve the generalization capability. This thesis examines the necessary measures to detect face presentation attacks in a more robust and generalised manner.

Keywords: Presentation attacks, Face presentation attack detection, deep learning, fusion, generation.

Acknowledgements

All praise and credit for this success go to Allah (SWT), as I acknowledge His never-ending blessings and benevolence in my life. I sincerely thank my principal advisor, Prof. Eyad Elyan, for his support and mentorship throughout my PhD. My most profound appreciation goes to Dr Pamela Johnston for her patience and guidance in making this thesis successful. I acknowledge Dr Adamu Ali-Gombe for providing valuable guidance in this research.

I am immensely thankful to my family for their unwavering love and support. I extend my heartfelt gratitude to my husband Aslam and my boys Ahmad, Ameer, and Adam, for their unconditional love, encouragement, tolerance, and empathy throughout this journey. Additionally, I would like to express my appreciation to my mother Ramlath, who gave me a sense of parental pride and assisted in taking care of my children. I recognize that this accomplishment would not have been feasible without all of you.

Abbreviations

ACER	Average Classification Error Rate
APCER	Attack Presentation Classification Error Rate
BCN	Bilateral Convolutional Network
BPCER	Bona fide Presentation Classification Error Rate
CDCN	Central Difference Convolutional Network
CNN	Convolutional Neural Networks
DMD	Dynamic Mode Decomposition
DNN	Deep Neural Networks
DoG	Difference of Gaussian
DSU	Domain Specific Unit
EER	Equal Error Rate
FAR	False Acceptance Rate
FAS	Face Anti-Spoofing
FFT	Fast Fourier Transform
FN	False Negative
FP	False Positive
FPAD	Face Presentation Attacks Detection
FR	Face Recognition
FRR	False Rejection Rate
FSL	Few-shot learning
HOG	Histogram of Oriented Gradient descriptors
HTER	Half Total Error Rate
IMQ	Image Quality Measure
LBP	Local Binary Patterns
LSTM	Long Short-Term Memory
MAFM	Multi-scale Attention Fusion Module
MLP	Multi-Layer Perceptron
NAS	Neural Architecture Search
NIR	Near Infra-Red
PA	Presentation Attacks
RF	Random Forest
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristic
rPPG	Remote Photo Plethysmography
SA	Simulated Annealing
SURF	Speeded-Up Robust Features
SVM	Support Vector Machine
SWIR	Short Wave Infra-Red
VIS	Visible Spectrum
ViT	Vision Transformer

Summary of publications

As part of the work in this thesis, the author published five articles. They were:

Journals

- Abdullakutty F, Elyan E, Johnston P. A review of state-of-the-art in Face Presentation Attack Detection: From early development to advanced deep learning and multi-modal fusion methods. Information fusion(pp. 55-69). Volume 75, November 2021. https://doi.org/10.1016/j.inffus.2021.04.015.
- Abdullakutty F, Elyan E, Johnston P, Ali-Gombe A. Deep Transfer Learning on the Aggregated Dataset for Face Presentation Attack Detection. Cognitive Computation (pp.2223-2233). Volume 14, November 2022. https://doi.org/10.1007/s12559-022-10037-z.
- Abdullakutty F, Johnston P, Elyan E. Fusion Methods for Face Presentation Attack Detection. Sensors. Volume 22, July 2022. https://doi.org/10.3390/s22145196.

Conferences

- Abdullakutty F, Elyan E, Johnston P. Face spoof detection: an experimental framework. In Proceedings of the 22nd Engineering Applications of Neural Networks Conference (EANN) (pp. 293-304). Cham: Springer International Publishing. 2021. https://doi.org/10.1007/978-3-030-80568-5_25.
- Abdullakutty F, Elyan E, Johnston P. Unmasking the Imposters: Task-specific feature learning for face presentation attack detection. In International Joint Conference on Neural Networks (IJCNN). IEEE. 2023. https://doi.org/10.1109/IJCNN54540.2023.10191953.

Declaration

I confirm that the work contained in this PhD project report has been composed solely by myself and has not been accepted in any previous application for a degree. All sources of information have been specifically acknowledged and all verbatim extracts are distinguished by quotation marks.

Signed Faseela .... Date: October 2023
Faseela Chakkalakkal Abdullakutty

Chapter 1

Introduction

The need to safeguard against identity theft poses a significant challenge for automated personal authentication that relies on Face Recognition (FR). The FR system, despite its highly accurate recognition results, is still susceptible to vulnerabilities. Imposters circumvent the FR system by presenting a photo, video, or mask that represents the identity of the real user. The term "Presentation Attack (PA)" refers to these types of attacks [1]. The Internet and surveillance video footprints serve as the source of facial images of individuals [2]. Thus, facial images are readily available and can be used to create PAs. As a consequence, FR systems have increasingly become targets of identity theft using PAs [3]. There is a profound negative impact of PAs on the reliability of FR systems, which are known for their non-intrusive, user-friendly, and cost-effective nature. Therefore, through the application of deep learning, this research strives to create novel techniques for Face Presentation Attack Detection (FPAD) that can effectively identify and prevent identity theft, while also enhancing their capacity to generalise to new and diverse contexts.

1.1 Background

In recent years, there has been remarkable progress in face recognition. There are, however, increasing attempts by individuals to deceive or manipulate security applications based on this widely used biometric modality. However, it is worth mentioning that even FR is vulnerable to a wide variety of attacks, which reduce its reliability [4]. Presentation Attacks (PA) are a common form of attack against FR systems. PAs are direct attacks on FR systems among direct and indirect attacks. A PA is carried out at the sensor level [5]. Imposters use pictures, videos, and masks to mimic the facial features of genuine users in order to circumvent the security of the face recognition system. Imposters do not require any knowledge of the system since these attacks are conducted in front of a camera. The ease of executing the attack, coupled with the availability of facial images and videos on the internet and through social media, makes presentation attacks the most common attack against the face recognition system [6].

PAs are the replica of authorised user faces in the form of photos, videos or masks. The photos can be printed on different materials or displayed on digital screens. Similarly, video can be displayed on screens of different resolutions. Masks, which are the most sophisticated form of PAs, are made from different materials [7]. Thus varying and emerging spoofing medium makes it harder to identify between real and fake facial images. Since PAs include images and videos of users, they have distortions caused due to the recapturing of which they have undergone. These distortions act as the cues to identify PAs from real images and carry out FPAD [8].

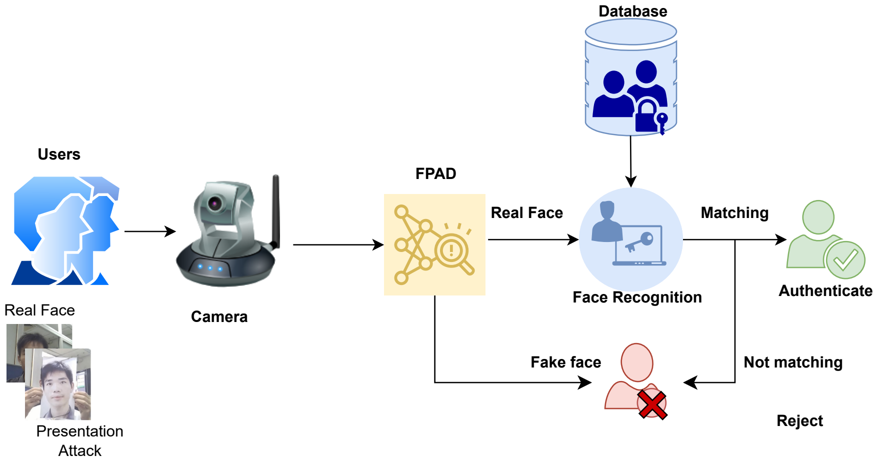


Figure 1.1: An FR system with face presentation attack detection (FPAD).

A face recognition system identifies the person matching the captured image with the user database. In the absence of an FPAD module, the FR system does not verify the genuineness of the captured image [9]. This trend reduces the secure authentication provided by FR. Hence, it is vital in the FR system to check the authenticity of the captured image before matching it with the user database. The FPAD module checks whether the captured image is genuine or not in FR systems. Fig. 1.1 illustrated the authentication procedure using an FR system with FPAD. The FPAD module verifies the captured images for genuineness. Only real images are then sent to the FR matching and authentication procedure. If the image is detected as PA, access is denied. Mismatched real images are rejected as well.

As presentation attacks are the common way of fooling face recognition, FPAD has gained serious attention among the biometric community. A plethora of state-of-the-art (SOTA) methods have been implemented using machine learning and deep learning [1]. However, these SOTA have deteriorated performance while using them in real-life scenarios. The primary reason is that earlier methods used available public datasets. These datasets were recorded in limited and often controlled settings with a very less number of subjects while including any one type of attack variant [10]. Hence the resulting methods have limited detection capacity while tested with different datasets or in real-life applications. Not only that, the detection performance even varies if the spoofing materials are different. Along with dataset variation [11], the method also should be capable of extracting suitable features before classification [12] to generalise against different variants of attacks.

1.2 Motivation

Presentation attacks can undermine the security of FR systems and negatively impact the authentication process. Manufacturing and technological advancements have made it possible to create novel PA variants [7]. The new PA variants are unseen attacks against the existing FPAD methods. The current FPAD methods, however, have already been trained on public Face Anti-Spoofing (FAS) datasets with limited variance in terms of size, attack type, resolution, and settings. It is essential that an FPAD method possess generalisation against various PA variants in order to have the best chance of detecting new and emerging PA variants [13]. Because of the limited generalisation capability of the current methods, new variants cannot be detected. In order to improve the generalisation of current data-driven methods, a training dataset with possible attack variants can be useful. Nevertheless, FPAD does not currently possess such comprehensive datasets. But, in some cases, data aggregation within the source domain can increase variance and thus enhance generalisation [14].

An FPAD verifies the authenticity of a facial image captured by an FR sensor. In light of this, it is highly relevant to extract and process corresponding features in order to improve PA detection and generalization [15]. The FPAD scenario also requires a tailored method, in addition to variance in the training dataset. There has been a recent emergence of hybrid methods in classification which combine different types of extracted features [16]. In addition, feature fusion from different pre-trained models can provide a potential solution for generalised FPAD. By extracting appropriate features using deep learning models, it may be possible to improve generalisation while detecting presentation attacks [17].

1.3 Research Objectives

This thesis has the following objectives:

- Critically review the current state-of-the-art methods in face presentation attack detection. Analyse publicly available datasets, attack types, and the associated challenges.
- Create an experimental framework to detect presentation attacks using a public dataset. Alter the training set variance by using different custom dataset partitions which were used by the state-of-the-art methods, including the official dataset partition. Carry out an impact analysis of custom dataset partitioning on presentation attack detection performance.
- Using data aggregation and deep transfer learning, devise a face presentation attack detection method using various pre-trained models. Assess the cross-dataset performance of the transfer learning models trained on the aggregated dataset empirically in the FPAD context.
- Utilise a hybrid fusion approach that combines both deep features and colour texture features to detect face presentation attacks effectively. Analyse the detection performance of the hybrid fusion methods using three different public FAS datasets with the corresponding deep transfer learning models.
- Implement a method for detecting face presentation attacks that utilises task-specific learning. Conduct a cross-dataset assessment of the proposed method to evaluate its generalisation ability across different datasets.

1.4 Contributions

This thesis makes the following contributions in the area of face presentation attack detection (FPAD):

- A comprehensive overview of the state-of-the-art methods for detecting face presentation attacks, a review of available publicly available datasets, evaluation metrics and a discussion of the challenges and possible future directions including data aggregation [11], hybrid fusion [12] and task-specific feature learning.
- A research that demonstrates the integration between data aggregation and deep transfer learning for PA detection. It demonstrates that to improve face presentation attack detection performance, it is essential not only to employ source domain aggregation but also tailored methods [11].

- A study that concludes, face presentation attack detection becomes more effective when features from different extraction methods are combined, thus reducing false positives [12].
- A novel method to learn task-specific features for face presentation attack detection in order to improve generalisation.

1.5 Thesis Structure

Chapter 2 presents a review of the necessary background for this research. This chapter discusses attacks on FR systems and the general taxonomy of detection methods. Additionally, existing Face Anti-Spoofing (FAS) datasets are reviewed in terms of their variance. A discussion of the latest trends in the detection of PAs is presented in this chapter. The chapter also discusses challenges and future directions pertaining to better generalisation against unseen attacks.

The impact of custom dataset partitions on the detection of face presentation attacks has been discussed in chapter 3. With the NUAA dataset [10], this chapter investigates how train-test partitions and variance in training data affect model performance.

Chapter 4 presents an experimental framework to detect face presentation attacks using data aggregation and deep transfer learning. Using popular public FAS datasets, an aggregated dataset was formed. pre-trained deep models trained on the aggregated dataset were assessed for generalisation capability.

In chapter 5, a hybrid fusion method combining deep and colour texture features is presented. The performance improvement in FPAD, using this fusion method, is evaluated using three FAS public datasets and pre-trained deep models. The chapter also includes a comparison of computational speed between the baseline method and the fusion method.

A task-specific feature learning procedure using deep pre-trained models is presented in chapter 6. Three pre-trained deep models were trained and evaluated using three public FAS datasets. In addition, a cross-dataset evaluation is conducted in order to assess the generalisation capability. To form fusion models, deep features are combined with various features such as texture, image quality, and frequency. A comparison of the performance of these fusion models is made with that of deep models that have been fine-tuned for task-specific learning for FPAD and its generalisation.

This thesis concludes with chapter 7 which offers a summary of the work and acknowledges its limitations. It also outlines potential future directions for further exploration.

Chapter 2

Research Background

Analysis of the current state of the art is critical prior to the development of novel, generalist FPAD methods. In this chapter, a background in the field of face presentation attack detection is presented. Presentation attacks (PA) are classified into several types and this chapter discusses their characteristics. With the advent of deep learning methods, effective feature learning was achieved in a variety of applications, including Face Presentation Attack Detection (FPAD). Moreover, deep learning methods outperform traditional methods for detecting presentation attacks. Furthermore, PAs have been detected by hybrid methods too. The purpose of this chapter is to provide an overview of current FPAD methods. A review of existing research is also conducted in order to identify the challenges and potential directions to be pursued in the field of FPAD. The main findings of this chapter have been published in "Information Fusion" in November 2021, [1].

2.1 Introduction

The primary objective of Face Presentation Attack Detection (FPAD) is to distinguish between a genuine and forged face in an image or video. As a result of analyzing the features extracted from the detected faces, the FPAD module identifies whether the PA is present or not. The authentication process is based on the output of the FPAD module and face matching. While traditional methods used machine Learning for PA detection, deep learning has been employed for this purpose in the last few years extensively.

Machine learning is a form of artificial intelligence that mimics the human brain's capacity to gain knowledge and comprehension through experience and enhancement without explicit programming. It encompasses algorithms or techniques that extract patterns from data and draw inferences based on those patterns. The fundamental goal of machine learning is to enable computers to learn independently, without human intervention or guidance, and adapt their actions accordingly.

Figure. 2.1 shows an example of a generic machine learning algorithm. A PA image is provided as input. Features are extracted from the image data and passed to the machine learning model to identify whether it is a real face or a fake face in the input image. ML models are used to detect the presence of PAs in images by analyzing the extracted feature patterns. The process of learning from data is referred to as training in machine learning. It is possible to train an ML model in three different ways; supervised learning, unsupervised learning, and semi-supervised learning.

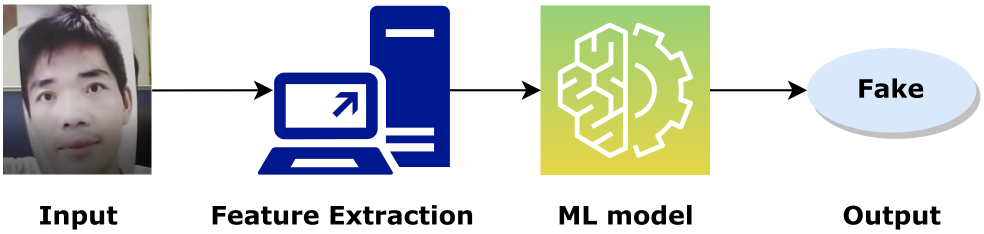


Figure 2.1: Working of a ML system.

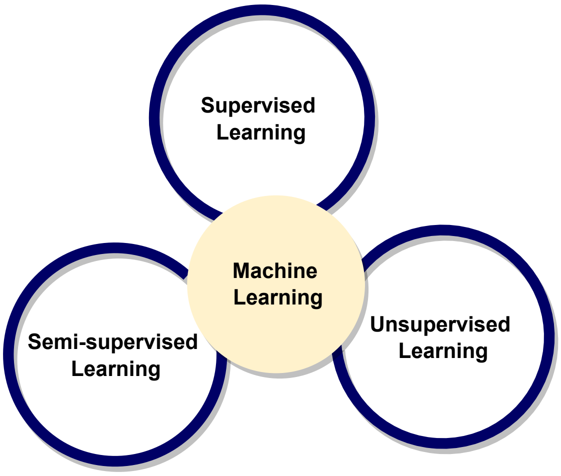


Figure 2.2: Learning and classification approaches used in ML.

Training is conducted using the labelled dataset in supervised learning. A prediction function is inferred by analyzing the training dataset. A sufficient amount of training will enable the system to provide targets for any new input. The model can also be modified accordingly if errors are found by comparing its output with the intended, correct output. An unsupervised machine learning approach involves algorithms that are trained on data that has not been labelled. Analyzing data sets in search of meaningful connections is the goal of the algorithm. In semi-supervised machine learning, supervised and unsupervised learning are combined to achieve the desired results. Data scientists feed this algorithm largely labelled training data, but it is left up to the algorithm to explore and analyze the data on its own. A supervised learning approach is the most commonly used training method in machine learning [18]. In the areas of classification, regression modelling, and ensemble learning, supervised machine learning models are useful.

Supervised machine learning generally uses two types of classifiers: probabilistic and linear. Mixture models are used in probabilistic classifiers. Each class is considered to be an element of the mixture. Mixture elements, which are generic models, provide the possibility of sampling a specific term. These classifiers are also referred to as generative classifiers. The Naive Bayes (NB), Bayesian network (BN) and maximum entropy classifiers are examples of probabilistic classifiers. There is also a type of supervised machine learning classifier known as a linear classifier. In linear classifiers, items with similar features are grouped together. It is the linear combination of these features that drives the classification decision in linear classifiers. Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), Logistic Regression (LR), Decision Tree (DT), Random Forest (RFs), and Neural Networks (NNs) are examples of supervised linear machine learning classifiers. Figure. 2.2 shows the learning methods and classifiers used in ML. For image classification, SVM and RF are widely used. In essence, FPAD involves the classification of images. The use of SVM and RF in FPAD using hand-crafted features has been extensively demonstrated.

Deep Learning (DL) is a sub-field of machine learning that uses Artificial Neural Networks (ANN) as its basic building blocks. ANNs have a structure similar to that of human neurons and process information in the same way as human neurons. In this manner, a deep learning model is able to carry out a variety of tasks in the same manner as a human brain and even more accurately. In recent years, deep learning models have been enhanced by improvements in computing performance and the availability of datasets. In contrast to traditional machine learning models, deep learning models operate on raw data.

The deep learning model is a neural network model with several layers and parameters between output and input. Figure. 2.3 illustrates a simple neural network model. As a general rule, neural networks are composed of three layers: the input layer, the hidden layer, and the output layer. As deep learning models use neural networks with more hidden layers, they are commonly referred to as deep neural networks. Deep refers to the fact that there are more hidden layers between the input and output layers. An ordinary neural network model may contain two or three hidden layers, whereas a deep neural network model or a deep learning model may contain 100 or more hidden layers.

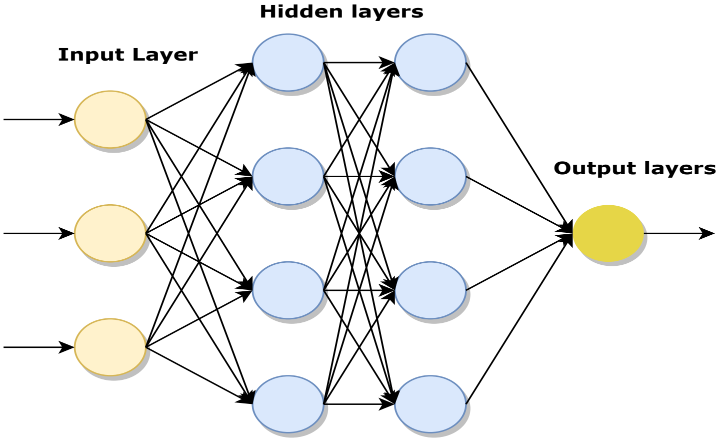


Figure 2.3: A neural network model.

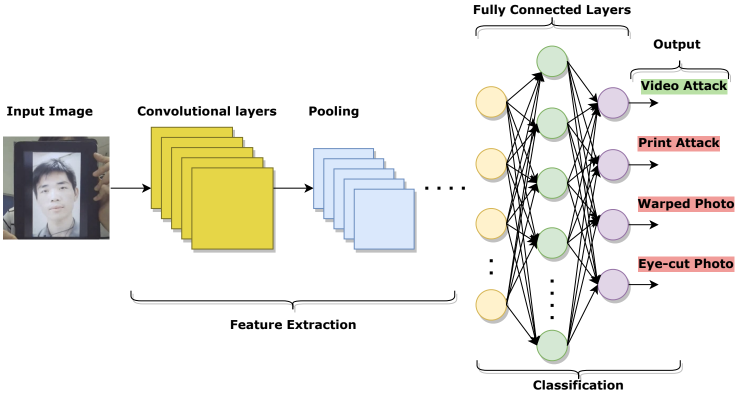


Figure 2.4: Generic representation of PA detection using Deep Learning.

The DL process provides automatic feature learning through the use of different layers. In DL, representations are learned hierarchically at different levels from the input data. As a result of the inherent feature learning capability of DL models, these models are more robust than traditional machine learning models. The architecture of a DL model consists of modules for extracting features as well as modules for classifying them. Figure. 2.4 illustrates how a deep learning model is used to perform PA detection.

Deep learning has different techniques including Convolutional Neural Networks

(CNN), Recurrent Neural Networks (RNN), Restricted Boltzmann Machines (RBM), Auto encoders and Extreme Learning [19]. CNN are the most commonly used model for various tasks. A CNN model normally consists of three types of layers, convolutional layers, pooling layers and fully connected layers as in Figure. 2.4. The convolutional layers, use different kernels or filters to generate feature maps at different levels from the input images. Pooling layers facilitate dimensionality reduction (width × height) of the input for the upcoming convolutional layers. The pooling strategies that are commonly employed include average pooling and max pooling. The 2D feature maps obtained from convolutional and pooling layers are transformed into a 1D feature vector through fully connected layers. The resultant feature vector can be directed towards either a classifier layer or for subsequent processing.

Transfer learning, a technique used in deep learning, allows models trained for a specific task to be applied to different datasets or tasks. There are two common approaches to transfer learning. Firstly, a pre-trained model can be utilized as a pre-built feature extractor for a similar task. Secondly, the pre-trained models can be partially or entirely fine-tuned by training them on a specific dataset to perform the desired task. Transfer learning is frequently employed in scenarios where data sizes are limited to avoid overfitting. One of the key benefits of transfer learning is the significant reduction in training time and computational resources. Notably, some widely used pre-trained image classification models are ResNet-50 [20], VGG-16 [21], and Inception V3 [22], which were trained on the ImageNet dataset [23].

While earlier methods for face presentation attack detection relied on machine learning with extracted features (section. 2.5), recent studies have leveraged the benefits of deep learning, particularly transfer learning, to detect presentation attacks (PAs) (section. 2.6). In more recent approaches, a combination of both handcrafted and deep features has been employed using hybrid fusion methods (section. 2.7).

2.2 Attacks on Face Recognition Systems

Attacks on Face Recognition (FR) systems are generally classified into direct and indirect attacks. Direct attacks occur at the sensor stage by presenting forged facial artefacts. Direct attacks are easier for the attacker since they require less knowledge about the attacked FR system. In the absence of appropriate protection, the system is highly susceptible to direct attacks. Indirect attacks affect pre-processing, feature extraction, database, matching, and decision modules. The attacker should be aware of the system to successfully execute these attacks [5]. By improving FR infrastructure, communication channels, and perimeter security, cyber security plays a direct role in preventing indirect attacks. Direct and indirect attacks influence the FR system as shown in Figure. 2.5. Duplicating and introducing facial artefacts to the FR system has become easier with technological improvements [6, 5]. Common direct attacks are:

- Presentation Attacks
- Disguise/makeup
- Modifications done through plastic surgery

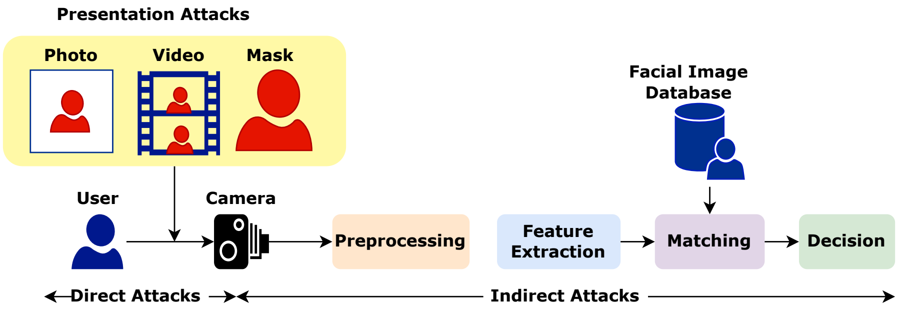


Figure 2.5: Attacks on Face Recognition system

Disguised faces are one type of direct attack. Disguise accessories can intentionally or unintentionally impersonate or obfuscate. Unintentional disguises include sunglasses, hats, or scarves. FR is vulnerable to various types of intentional and unintentional disguise accessories. The authors of [24] observed that the facial portions under disguise accessories provide false data and FR cannot use these for identifying a user. Hence disguise accessories facilitated the hiding or imitating of identity. These types of disguise attacks are prominent in border crossing and airport security applications [25].

Makeup is another direct attack similar to disguised faces. It is harder to identify makeup attacks as they have a close resemblance to the real face [26]. While keeping the genuine appearance of the human face, makeup can easily obfuscate the true identity of the user. Among the direct attacks, it is easily available, cheaper and variable in nature.

Plastic surgery is a direct attack, too. Face regions including nose, eyes, lips, ear or bone structure are reformed to obtain desired appearances. These cause long-lasting changes in features in specific facial regions. The reference database may contain the pre-surgery sample for face recognition. In this case post-surgery biometric recognition becomes challenging due to the alterations [27]. Some disease treatment surgeries can also unintentionally increase variations in facial appearance [28].

Attackers make use of eyeglasses, facial hair and caps either to impersonate or obfuscate. Such effects are generated using adversarial methods too. These adversarial-generated attacks are able to mislead the classifier in deep learning-based FR systems [29]. Slight perturbations added to the image can also lead to misclassification in FR systems. These perturbations are physically imperceptible or even ignored by human eyes, yet capable of causing misclassification in FR systems [30]. Thus, synthetic images generated through adversarial methods and modified images with adversarial perturbation act as PAs [31]. Attackers add perturbations in two ways: ‘no target’ and ‘dodging’. In ‘no target’, the aim is to hide the identity of the user, whereas in ‘dodging’, perturbation is added to access the identity of a target user. The authors of [32] introduced an eyeglass printing method to generate physically realisable attacks. Sharif et al. [33], using 3D-printed eyeglasses, generated attacks to execute impersonation. Using an infrared lighting cap [34] was able to create adversarial physical attacks. Adjusting the positions, size and intensity of the infrared dots generated by this cap, the attacker could pass through the security system. Nguyen et al. [35] proposed a more convenient method to create adversarial attacks using light projections. Real-time physical attacks were created by changing the camera-projector setting suitable to the attacking environment.

2.3 Presentation Attacks (PA)

Presentation attacks (PA) [36] are used either to impersonate or to obfuscate a user while passing through an FR system. Impersonation is carried out by copying a genuine user’s facial attributes to gain access through FR systems. Obfuscation is used to hide the user’s identity using various methods such as glasses, makeup, disguised face and facial hair [37]. A generic FR system detects faces from the image or the video input and recognises authorised users with respect to the reference database. PAs have duplicate facial features in the form of photos, videos or masks. This will assist the attacker to invade the security system if the FR does not have a detection module to differentiate between genuine and fake faces. Hence, PAs affect the proficiency of FR systems in security applications [9].

PAs are broadly classified into 2D and 3D attacks as can be seen in Fig. 2.6. Photo attacks and replay attacks are 2D attacks [38], whereas mask attacks are included in 3D attacks [7]. 2D attacks are very common and are carried out by presenting facial artefacts using photo or video to the sensor [5].

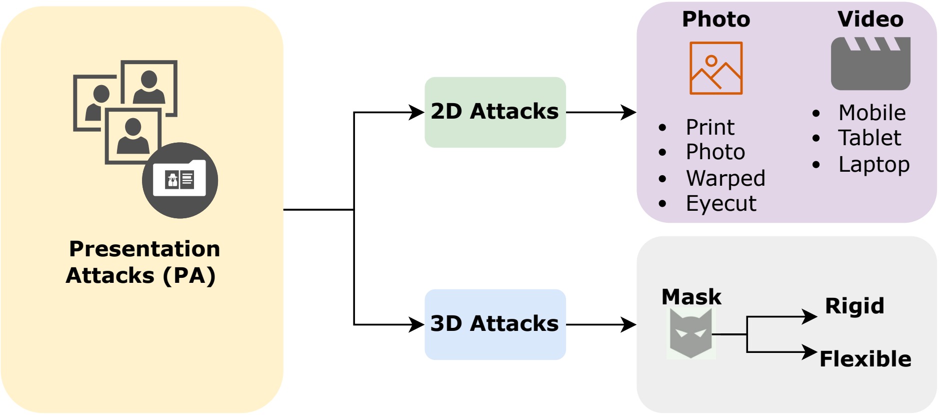
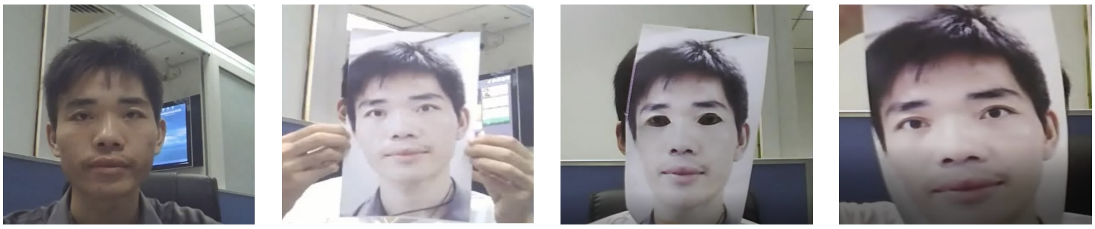


Figure 2.6: Types of Presentation Attacks

2.3.1 Photo Attacks

Photo attacks are executed by displaying a photograph of the genuine user to the FR system camera. Flat printed photos, digital display of photos, eye-cut photos, and warped photos are all variants [39]. It is possible to print colour images of genuine users with great ease and at a low cost. Photos can be displayed on digital devices with high-resolution screens. Furthermore, social media gives easy access to authentic facial images. As a result of technological advancements in digital cameras, high-quality photos can be obtained by using hidden cameras. Consequently, print attacks are more common because they are easier to execute.

In cut-photo attacks, there will be holes in the position of the eyes and mouth. These holes help the imposter to imitate live features like eye blinking and mouth movements [39]. Spoofing an FR system which works based on the liveness of the user, can be carried out with these types of photos. Cut-photo attacks are harder to detect compared to flat printed photo attacks [40]. Figure. 2.7 gives an example of authentic and photo attack images from CASIA FAS dataset[41].

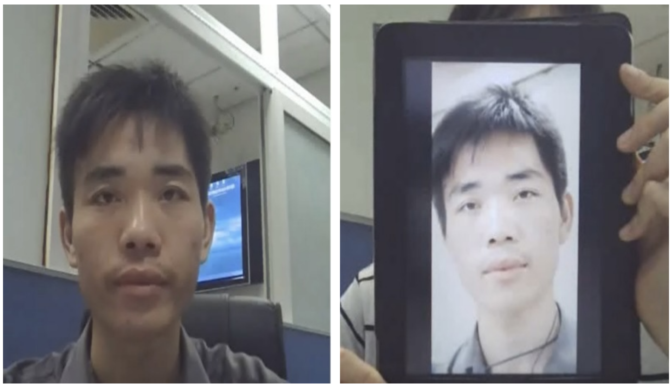


(a) Bona-fide (b) Print Attack (c) Eye-cut photo (d) Warped photo

Figure 2.7: Figure showing bonafide and photo attack variants [41]

2.3.2 Video Attacks

Video attacks, also known as replay attacks in face presentation attack detection literature, are more sophisticated forms of attacks. Videos of genuine users are presented to the FR system [39, 5], using a mobile, tablet or any other digital devices [6, 42]. Since the video consists of both movement and background information, distinguishing fake users from bona fide users in these cases is a challenge[43].



(a) Bona-fide (b) Video Attack

Figure 2.8: Figure showing bonafide and video attack [41]

Video attacks can emulate liveness, unlike photo attacks. Thus, FR systems that are susceptible to photo attacks will be even more vulnerable to video attacks. Video samples of genuine users can also be found on social media.

2.3.3 Masks

While FR sensors capture images for authentication, the imposter can wear a mask with the features of a genuine user. 3D masks possess face-like depth and this is a unique challenge in detecting the 3D mask PA. Ramachandra et al. [44] experimented to find vulnerabilities in two commercial FR systems and observed that, due to the False Acceptance Rate (FAR) threshold, these systems were vulnerable to customised silicon masks. In a similar study, Bhattacharjee et al. [45] investigated customised silicon masks and found that FR systems were highly vulnerable to flexible mask attacks.

There are various types of masks made of distinct materials. Much of the literature covers 2D-type attacks as, historically, it was difficult and costlier to produce 3D masks. Lately, there have been developments in 3D printing technologies which have provided cheaper and easier ways to produce 3D masks [7]. 3D masks are made of different materials [27]. Hard or rigid masks can be made from paper, resin or plastic. These masks are used as an improved variant of photo attacks. These cheaper types of masks appear visually similar to real faces due to the enhanced printing options available nowadays. Masks which are produced using silicon or latex are soft, and flexible and adapt to different facial shapes and sizes. They have close similarities with genuine facial texture and colour. It makes these soft masks more challenging to detect than rigid masks.

2.4 Face Presentation Attack Detection (FPAD)

Face recognition systems detect faces from images or videos captured by cameras. Following the feature extraction, the captured face is compared with the enrolled faces in the reference facial database. Authentication is provided if the face matches one of the faces enrolled in the database, as shown in Figure. 2.9. The FR system examines the authenticity of the user rather than the authenticity of the captured image or video.

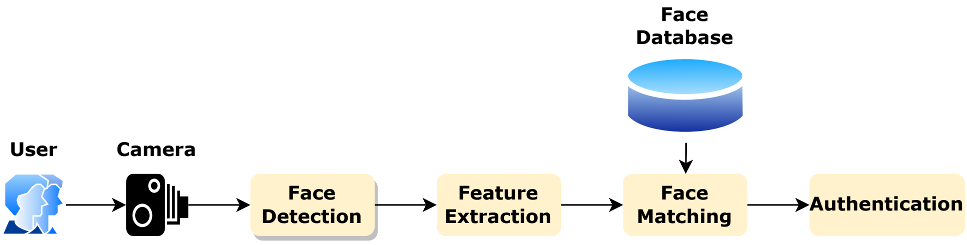


Figure 2.9: A generic Face Recognition (FR) system

The primary objective of Face Presentation Attack Detection (FPAD) is to distinguish between a genuine and forged face in an image or video. As a result of analyzing the features extracted from the detected faces, the FPAD module identifies whether the PA is present or not. The authentication process is based on the output of the FPAD module and face matching, as shown in Figure. 2.10.

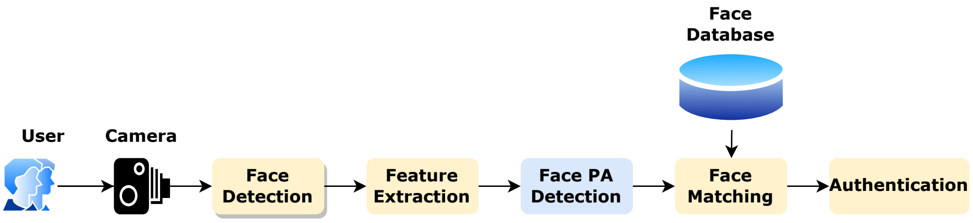


Figure 2.10: FR system with face presentation attack detection

Face recognition systems process static and dynamic cues using different techniques for spoof detection. Earlier feature-based methods mainly utilised hand-crafted feature extraction, using machine learning to detect presentation attacks. Local Binary Patterns (LBP) [46, 47], Histogram of Oriented Gradient descriptors (HOG) [48, 49], Speeded-Up Robust Features (SURF) [50], Difference of Gaussian (DoG) [51, 10] were the techniques adopted in hand-crafted feature methods. The emergence of deep learning methods provided effective feature learning in many applications. Moreover, deep learning methods provided better detection performance compared to hand-crafted methods.

2.5 Hand-crafted feature methods

Handcrafted feature methods extracted various features from the face images to distinguish between real and fake faces [5]. FPAD involves processing features extracted from the captured face images followed by classification [52, 53, 36]. Texture, temporal data, image quality and life signs are typical features processed to identify PA. Feature-based methods are classified into two types: static and dynamic [5]. Texture and image quality-based PAD methods are examples of static approaches, whereas temporal (or motion-based) and vital signals based methods are dynamic approaches.

Static approaches include texture and image quality-based techniques. They do not rely on temporal information and a single image is processed at a time to detect spoofing [42, 28]. By processing each frame independently, static approaches can perform anti-spoofing tasks using video. The processed outcome of the majority of frames is taken into account to form the final decision. Due to their performance, low computation and low cost, static approaches are popular. In comparison with dynamic approaches, static approaches are faster [5].

Through micro-textural analysis of the facial image, textural PAD methods distinguish real images from fake ones [42]. These methods identify the photo and replay attacks [5]. Local Binary Pattern (LBP) descriptors are the most widely used technique in texture-based-PAD methods. Authors of [47] proposed a PAD method using LBP for textural analysis for photo attacks. Replay attack detection was explored in [46] using the same technique tuned for video attacks. The advantages of these methods were easy deployment and no user interaction. However, these methods required feature vectors and exhibited poor performance with low-resolution images [54].

Presentation attacks affect the quality of the image [42]. Spoofing images are prone to distortions like surface reflection, Moiré-effect, colour distortion, and shape deformation [55]. In [43], the authors detailed the various distortions an image may be subject to due to spoofing medium, camera and printing. Spoofing medium (LCD or paper) causes specular reflection. Blur is introduced if the camera is out of focus while capturing the spoofing image. Reduced resolution of printed paper or LCD can also create colour distortion. Spoofing mediums add noise to the image [56]. The frequency histogram of a spoof image would be different to that of a genuine image. Face PAD systems use these quality variations in the image as cues while performing spoof detection.

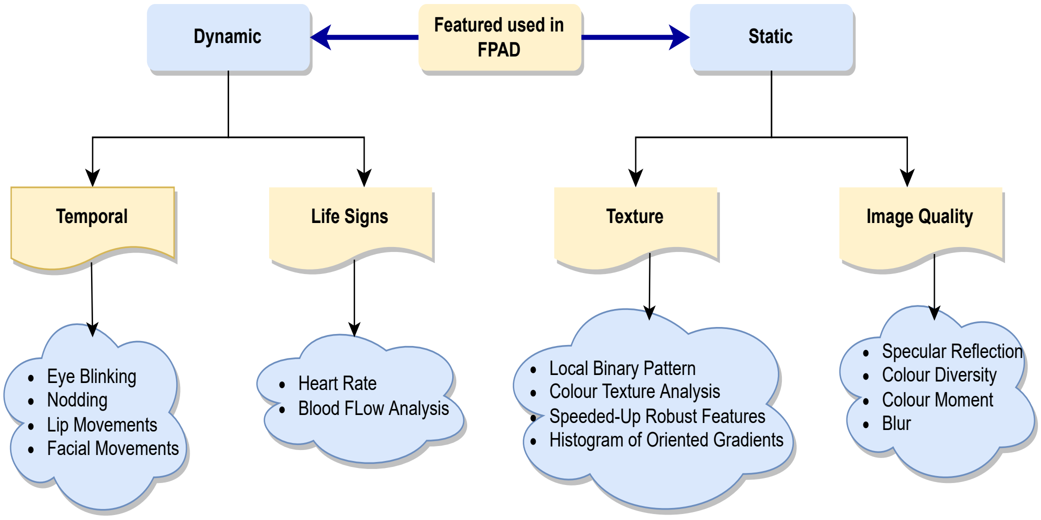


Figure 2.11: Features, which are used to detect PAs

Dynamic approaches depend upon temporal information to identify the presence of spoofing in FR systems [57]. They process life signs or motions to verify the liveness of the input presented to the facial sensor in the FR system. In dynamic approaches, performing temporal feature analysis, based on relative motion in the video provides information for spoof detection. Hence, dynamic approaches require more computational time compared to static approaches [5]. Some dynamic approaches rely on life signs too. Pulse, eye blinking, lip movement, and head rotation can be used to confirm the liveness in the FR system [42].

A temporal information-based algorithm Dynamic Mode Decomposition (DMD) was used in [58] to identify liveness. The authors used eye blinking and lip movements as motion cues. Motion-based PAD techniques demand user cooperation during the identifying process. This affects the processing time in FR system [42]. Some motion-based methods exploit impulsive movements of the facial parts in the input videos [59]. In [60], the authors followed a multiple-motion-cue-based method, considering eye-blinking, chin and lip movement. The authors of [61] presented liveness detection methods based on pupil tracking. Remote Photo Plethysmography (rPPG) is used for the acquisition of vital signals such as pulse or heart rate without contact with the human body. Since these vital signals are extracted from live faces, they act as the perfect cues for liveness. Face liveness detection methods presented by authors of [62] utilized pulse cues from videos. Pulse detection using rPPG was effective in 3D mask attack detection [63]. In [64], the authors presented a face liveness detection approach based on blood flow analysis. rPPG and patch CNN-based method was adopted in [65] to detect face liveness too.

2.6 Recent trends in detecting face presentation attacks

Deep learning-based methods have been successfully applied to various domains including speech enhancement and recognition [66], lip reading from visual content [67], analysing intractable and complex biological datasets [68], security and intrusion detection [69], and others. Convolutional neural networks in particular, have introduced remarkable developments in computer vision applications, especially in biometrics [70]. Deep learning, along with its inherent feature learning capability, constructed a novel path to solve the anti-spoofing challenge. Existing methods based on deep neural networks, show excellent intra-dataset performance. However, these methods have also exhibited poor cross-dataset performance and unseen attack detection [71, 72].

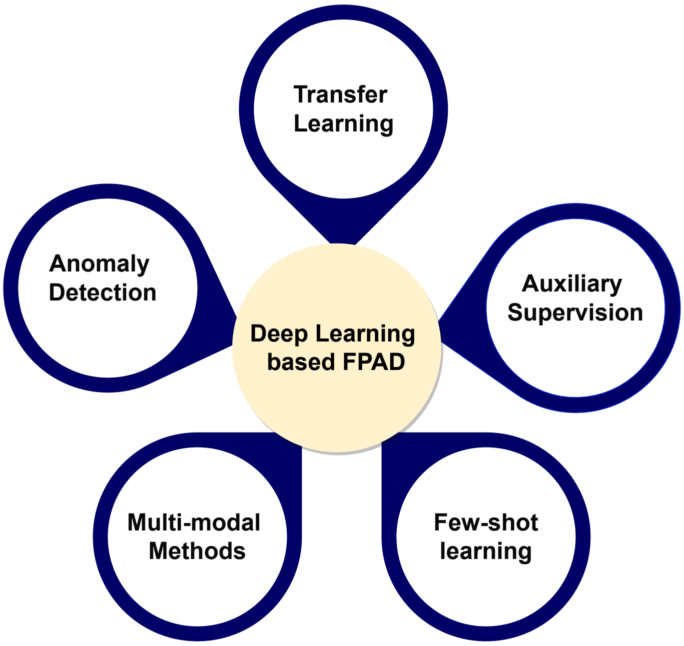


Figure 2.12: Recent trends in deep learning based FPAD

In the last few years, there has been a trend towards improving generalisation in PAD. In particular, unseen attack detection methods involve trying to accurately classify genuine samples and consider any sample except genuine ones as attacks. Some existing approaches have used only genuine samples for training so that the proper clustering and classification of genuine face would lead to desired detection of unseen attacks. These methods followed one class classification, as opposed to earlier models which followed binary classification for face PAD. A typology of recent trends in deep learning based FPAD is shown in Figure. 2.12.

2.6.1 Transfer Learning

Transfer learning is the process of re-utilising the learned features from a base network using base dataset to a target network to be trained with target dataset and task. Transfer learning helps to avoid overfitting when the training data is limited [73]. As training is not started from scratch, it also saves on computational resources. Nagpal et al. [74] analysed different CNN models for face anti-spoofing and their performance in detecting presentation attacks. Based on their research the authors recommended transfer learning with a deeper model utilizing lower learning rates for restricted computational resources. Lucena et al. implemented transfer learning for spoof detection in their work [75]. Among the deep learning methods deployed in FPAD, transfer learning is the most common one.

Domain adaptation and domain generalisation utilised a transfer learning approach to improve generalisation in FPAD. In domain adaptation, information from a source domain is transferred to a target domain using different techniques [76, 77]. Yu et al. [78] developed a neural architecture search-based face anti-spoofing (NAS-FAS) system. This method used central difference convolution and pooling. Transfer learning approach was applied on NAS for the spoof detection task. However, a cross-dataset evaluation for 3D mask attacks with NAS-FAS showed that challenges still remain in the generalisation capacity of even transfer learning approaches.

Domain adaptation

Domain adaptation mitigates the disparity between source and target domains. It facilitates feature learning in scenarios with limited training data. Hence, the generalisation capacity of face PAD can be improved using this method. In domain adaptation the model learns from the source domain on related distinct target domain [79]. Hence, recent research has utilised this technique to mitigate domain shift. Yang et al. [76] introduced domain adaptation in their research on personal specific face anti-spoofing approach. This approach transferred source domain subject-specific information on real and fake samples. This information facilitated the synthesis of virtual fake samples for subjects without fake samples in the target domain. Spoof detection was performed using a trained classifier for each person in this method. In real-life scenarios, there would be genuine samples without corresponding fake samples. However, personal-specific models required samples from all attack variants in the target domain to attain desired performance. This method also demanded more source domain fake samples for generalisation enhancement.

Inspired by the applications of Generative Adversarial Networks (GANs) [80] in various compute vision applications, Wang et al. [79] presented a domain adaptation method using them to address FPAD problem. Adversarial domain adaptation combined with deep metric learning assisted this model to outperform other state-of-the-art methods in both cross-dataset and intra-dataset evaluation. The authors extended this method using an unsupervised adversarial domain adaptation technique (UDA-Net) in [81]. UDA-Net carried out unsupervised adversarial domain adaptation. This facilitated extraction of common features associated with both target and source domains. As DR-Net assisted to transfer domain-independent information, it enabled better spoof detection in an unlabeled target domain. The authors carried out an extensive evaluation on more publicly available datasets.

Zhou et al. [82] adopted a multi-layer domain adaptation technique for spoof detection in face recognition systems. In order to reduce the disparity between source and target domains, the authors used a Multi-Layer Maximum Mean Discrepancy (ML-MMD). Similarly, Nikisins et al. [77] used domain adaptation by transferring facial features from RGB domain to multi-spectral domain. Domain adaptation was carried out using autoencoders. In this model, a set of multi-channel encoders were used for feature extraction. Classification of these features was performed by Multi-Layer Perceptron (MLP). The authors of [83] evaluated domain adaptation through domain-guided pruning of CNN. Recent domain adaptation research in face PAD is presented in Table. 2.1.

Table 2.1: Face PAD using domain adaptation method

Author	**Method**	**year**	**Datasets**
Yang et al. [76]	Person Specific Anti-spoofing	2015	CASIA, REPLAY ATTACK
Wang et al. [79]	Adversarial Domain Adaptation	2019	CASIA, MSU-MFSD, REPLAY ATTACK
Zhou et al. [82]	Multi-Layer Maximum Mean Discrepancy	2019	CASIA, REPLAY ATTACK
Nikisins et al. [77]	Multi-Channel Encoder	2019	WMCA
Mohammadi et al. [83]	Domain guided pruning of CNN	2020	REPLAY Mobile, SWAN, WMCA
Wang et al. [81]	Unsupervised Adversarial Domain Adaptation	2020	Idiap, MSU, CASIA, ROSE-YOUTU, CASIA-SURF, OULU

In domain adaptation, trained features are aligned to the target features to achieve better generalisation capacity through adapting the features of the target or test domain. However, for unseen attack cases, the target domain may be unknown and this would impact domain adaption.

Domain Generalisation

Domain generalisation is one of the techniques adopted by the biometric community to acquire generalisation in unseen attack scenarios. In existing face PAD methods, there is a bias towards the cues learned from training data. This impedes generalisation against unseen attacks with different environments, devices, lighting conditions or materials.

Costa-Pazo et al. [14] adopted domain generalisation for PAD. The authors designed a framework Generalised PAD (GPAD) to address the generalisation problem and suggested an aggregate dataset with variance in attacks, lighting, capture devices, and resolution. The GRAD-GPAD (Generalisation Representation over Aggregated Datasets for Generalised Presentation Attack Detection) provided a common evaluation method for face anti-spoofing techniques.

Saha et al. [84] addressed domain shift in face PAD using a domain-agnostic model. A class-conditional domain discriminator and gradient reversal layer were utilised to learn domain-independent features. Source domain features were learned through training using multiple datasets. The model showed improved generalised feature learning across multiple domains for print and video attacks. The multiple domains were formed due to the variations in illumination, background, printers, display screens, and the quality of recording devices. Wang et al. [85] utilised GANs to address unseen attack detection. The adversarial domain adaptation facilitated transferring source domain features to the target domain. This technique included Disentangled Representation learning (DR-net) and Multi Domain learning (MD-net). DR-net learned disentangled features. MD-net learned the generalised features across multiple domains using these disentangled features from these domains. Evaluation with CASIA, REPLAY -ATTACK, MSU and OULU-NPU datasets provided an improved cross-domain performance compared to existing state-of-the-art methods. However, the experiments also confirmed the fact that a larger dataset with more attack variants would be required for effective unseen attack detection.

Shao et al. [86] proposed another domain generalisation method without using the target domain data. Adaptive and automatic learning of generalised features was facilitated by a multi-adversarial deep domain generalisation module. Integrating a dual-force triplet-mining constraint enhanced the disparity in the generalised feature space. The model used auxiliary depth supervision to further improve generalisation. Unlike the aforementioned models, Jia et al. [87] followed a single domain generalisation. Retaining the boundary of feature domains of real and fake faces has become increasingly difficult due to novel attacks. In order to avoid grouping and extracting generalised features from multiple domains, the authors used a single domain. Use of Asymmetric Triplet Mining ensured the effective clustering of real face features while spreading away the fake ones. Zhang et al. [88] introduced an FPAD model disentangling features into live and content features. Depth supervision and translated images were utilised in this model.

The existing domain generalisation approach transferred generalised features from the source domain to a pre-defined distribution. However, this distribution might not be an optimal feature space [86]. Learning discriminant features requires multiple components in these models. Elimination of any of these feature discriminators might deteriorate the generalisation capability of the model [87].

2.6.2 Anomaly Detection

Unseen attack detection was addressed using an anomaly detection approach in recent research. Anomaly detection followed a one-class classification. From Table. 2.2, it is evident that this approach had gained more popularity in the last few years for unseen attack detection. In face PAD problem genuine or live face images are considered normal samples, whereas all possible attacks form the anomalous sample space. It has been found that the genuine class has lower variance within the feature distribution and forms a close cluster. They had more generalised features than the attacks. Attacks, on the other hand, can vary substantially from one another. The higher variance in attacks results in anomalies in the feature space. Using this close cluster behaviour of genuine samples in the feature space, anomaly detection techniques have classified authentic faces more accurately. Any samples outside the margin of the genuine sample cluster would be considered attacks. Since the real face sample has a defined class, unseen attacks can be detected.

Arashloo et al. [89] introduced anomaly detection for face PAD. This technique used only genuine face samples for training. The authors set up a new evaluation protocol to gauge the effects of unseen attacks. In terms of generalisation, these fake or negative samples represented all the spoofing samples. This method produced comparable results with the models using binary classification. In [90], Arashloo and Kittler proposed a similar technique to address unseen attack scenarios. The authors incorporated multiple kernel fusion, client-specific modelling, sparse regularisation and probabilistic modelling of score distributions to enhance the performance of the system. Through extensive evaluation using different datasets, it was shown that the method performed better than the existing state-of-the-art models in unseen attack detection.

Anomaly detection was explored by Nikisins et al. [91] too. Similar to [89], the authors used only genuine samples for training. Feature space was created using Image Quality Measures (IMQ) in this model. A Gaussian Mixture Model (GMM) to find out the probability distribution of genuine samples. Combining REPLAY-ATTACK, Replay-Mobile and MSU-MFSD public datasets, an aggregate dataset was formed. Compared to the binary classification methods, the designed model exhibited better generalisation when tested with the aggregate dataset. In a similar research, Fatemifar et al [92] used a client-specific model. In one class classification, each biometric trait has scores which would be distinct for genuine and attack samples. In this way, a threshold can be defined to distinguish between real and fake images. A client-specific threshold was set which provided a better distinctive capability to categorise genuine and attacks. This method exploited only real face information to implement a perfect anomaly detection approach. However, more mechanisms might be needed to refine single-class learners if the training data included fake samples. Fatemifar et al. [93] presented another subject-specific model. They fused the individual one-class classifier using a new normalisation technique in this ensemble learning method. A weighted average fusion strategy was used in the model.

Table 2.2: Anomaly detection approaches in recent face PAD research

Author	**Year**	**Remarks**
Arashloo et al. [89]	2017	A new evaluation protocol to detect the affects of unseen attacks
Arashloo and Kittler [90]	2018	Multiple kernel fusion, client-specific modelling, sparse regularisation, probabilistic modelling of score distributions
Nikisins et al. [91]	2018	Image Quality Measures (IQM), Gaussian Mixture Model (GMM)
Fatemifar et al. [92]	2019	Subject specific models
Perez-Cabo et al. [94]	2019	Deep Metric Learning
Fatemifar et al. [93]	2019	Client Specific Modeling
Abduh & Ivrissimtzis [95]	2020	Convolutional Autoencoder, in-the-wild training images
Li et al. [96]	2020	Hypersphere loss function
Feng et al. [97]	2020	A spoof cue generator and an auxiliary classifier.
Baweja et al. [98]	2020	Pseudo-negative class samples

Deep metric learning was used in anomaly detection to address generalisation. Perez-Cabo et al. [94] proposed this method and evaluation was carried out using GRAD-GPAD [14]. Metric learning-based loss provided lower intra-class variance and higher inter-class separability. Better classification of fake and genuine samples resulted using metric learning-based approach. Feng et al. [97], presented another anomaly detection-based face PAD. In this method, the framework had a spoof cue generator and an auxiliary classifier. The model used a residual learning network to extract the spoof cues. The method achieved good state-of-the-art performance in unseen attack detection. In [96], Li et al. addressed face PAD in an open setting with an anomaly detection method. They introduced a new hypersphere loss function for end-to-end learning. Real faces formed a close cluster near to the origin of the hypersphere sustaining intra-class compactness. The attack samples are scattered at a specific distance from the genuine face cluster in the feature space to maintain the predefined margin between real and fake features. Hypersphere loss identified these attacks directly without using a separate classifier. Baweja et al. [98] introduced a novel training approach for anomaly detection. The absence of negative samples made end-to-end learning in one class classification non-viable. Hence, the authors proposed a "Pseudo-negative class" sample feature space, which helped the model in learning better decision boundaries between genuine and fake samples. The pseudo-negative class was modelled using a Gaussian distribution. Unlike other existing OCC models, end-to-end learning was carried out for both classifier and feature representation. The authors of [95] included in-the-wild images in the training dataset of a one-class classifier. These images were recorded in an uncontrolled environment. Hence, features learned during training facilitated model operation in an uncontrolled environment. This enhanced unseen attack detection.

2.6.3 Few-shot and zero-shot learning

Few-shot learning (FSL) [99] is the process of learning from few samples with the supervised data. FSL is suitable to applications which require large scale data from supervision. FSL has only a small number of labelled target samples. When the number of these samples for target class is zero, FSL is called zero-shot learning. Since the requirement of target samples are very few or zero (in zero-shot scenario), FSL is suitable for detecting unseen or novel attacks. Recent research made use of this advantage of FSL to detect unseen attacks in face PAD.

Qin et al. [100] proposed face PAD using zero-shot and few-shot approaches. The authors designed Adaptive Inner-update Meta Face Anti-Spoofing (AIM-FAS) with meta-learning. Using pre-defined live and fake samples along with a few samples of unknown attacks, the model carried out spoof detection. The meta-learner provided better discrimination between live faces and attacks. With the adaptive inner update, the discriminative capacity is enhanced, improving generalisation. Liu et al. [101] used a zero-shot approach to address the unseen attack detection in face PAD. The authors used a deep tree network to learn the semantic attributes of pre-defined attacks in unsupervised methods. Even though live samples clustered well in the feature space, they positioned very close to a specific group of attacks like transparent masks, funny faces, obfuscation makeup and paper glasses. This made detection more challenging in such scenarios and implies that these attacks might be more challenging to detect.

2.6.4 Auxiliary Methods

Anti-spoofing is considered a binary classification problem. Hence, the majority of the anti-spoofing models follow binary supervision. Nevertheless, binary supervision has demerits too. Even though it provides arbitrary cues to detect spoofing, some of the spoof patterns may disappear over the feature duplication process. This results in poor generalisation [102]. To overcome poor generalisation, auxiliary supervision has been used in a number of recent researches. It has been shown that auxiliary supervision with end-to-end learning can provide better anti-spoofing [103]. The methods which use auxiliary data are presented in the Table. 2.3. As in the table, depth was used as an auxiliary feature in the majority of the existing models.

Table 2.3: PAD with auxiliary supervision

Method	**Auxiliary Cues**	**Attacks**
Patch CNN [103]	Depth	Print, Replay
CNN-RNN [102]	Depth, rPPG	Print, Replay
Frame-level CNN [104]	Pixel wise binary	Print, Replay
CNN with OFFB and ConvGR [105]	Depth	Print, Replay
Central Difference CNN [106]	Depth	Print, Replay
Multi-Spectral Central Difference CNN [107]	Pixel wise	Print
Bilateral Convolutional Network(BCN) [108]	Human material	Print,Replay
Bipartite Auxiliary Supervised Network (BASN) [109]	Bipartite (Depth, Reflection)	Print, Replay
Contextual Patch-Based CNN [65]	rPPG	2D, 3D
Patch CNN [110]	Depth	Print, Video
SLNet [111]	Disparity	Print, Video
Image Decomposition [56]	Noise	Photo

Atoum et al. [103] introduced the depth supervision for anti-spoofing by proposing a depth supervised patch-based CNN. From the random patches, local features are extracted. These features were fused with a depth map to identify spoofing. A similar auxiliary supervised approach using depth and Remote Photoplethysmography (rPPG) supervision was proposed by Liu et al. [102]. The authors used a CNN and Recurrent Neural Network (RNN) combination for spoof detection. rPPG facilitated temporal information extraction using the difference in live signals for live face and spoof images. Distinct from the above-mentioned single-frame PAD methods with auxiliary supervision [103, 102], a multi-frame approach was followed by Wang et al. in [105]. This approach exploited temporal information along with depth supervision. The authors followed the distinguishing patterns of temporal depth and motion between live and spoof images in the temporal domain. This approach facilitated efficient spoof detection under depth supervision by examining complex facial variations and motions.

Auxiliary methods used depth and temporal features for supervision. The acquisition and processing of these features might take a longer time. Nevertheless, in a real-time scenario, especially in mobile devices this delay would not be acceptable. As the depth calculation consumed more computational resources and time, George et al. [104] followed a pixel-wise supervision PAD method. This method was claimed as a suitable approach for mobile devices as it avoided the pixel-wise depth calculation. In order to extract more generalisable features in auxiliary supervised PAD, Kim et al. introduced a novel Bipartite Auxiliary Supervised Network (BASN) [109]. This approach used auxiliary cues from both live face and spoof images, distinct from existing PAD methods with auxiliary supervision.

Following the aforementioned auxiliary supervised methods and leveraging the Central Difference Convolution (CDC), Yu et al. [106] introduced a spoof detection approach using Central Difference Convolutional Network (CDCN). By utilising Neural Architecture Search (NAS) architecture, low, mid and high level features were extracted. These features were fused using a Multi-scale Attention Fusion Module (MAFM). CDC provided better results by combining intensity and gradient information. Apart from achieving generalisation in face pose, expression, spoof medium, cross or unknown attack variants, this approach showed considerable performance in terms of domain shift. The authors extended the methods incorporating multi-spectral mode in [107] using two fusion strategies for the modalities. The fusion was done either by input-level fusion via concatenating three-modal inputs to $256x256x9$ directly or score-level fusion via weighting the predicted score from each modality. Yu et al. [108] also proposed a human material recognition model for face spoof detection. The authors included a Bilateral Convolutional Network (BCN) for capturing human material patterns. The BCN was able to learn macro-micro features associated with the material. A multi-level feature refinement module along with multi-headed supervision facilitated enhanced BCN performance by refining multi-scale features and learning shared features.

Authors of [65] proposed a method incorporating rPPG and textural information to attain generalisation in terms of 2D and 3D mask attacks. Multi-scale long-term statistical spectral features for rPPG information was incorporated with contextual patch CNN. Remote Photoplethysmography (rPPG) provided 3D mask and photo attack detection while textural cues identified the replay attacks. Liu et al. [110] developed a face PAD combining Patch CNN and Depth-based CNN. This approach was designed as a PAD for mobile devices. Depth-based CNN showed degraded performance for low-resolution images, whereas Patch based CNN showed low performance for high-resolution images. The combination of these two improved the overall PAD performance in mobile devices. Unseen attack detection was addressed by learning disparity maps and training end-to-end classifiers simultaneously. Rehman et al. [111] proposed an approach similar to the depth-supervised auxiliary method. The learned disparity maps facilitated better detection of unseen attacks. Auxiliary supervision was investigated by Jourabloo et al. [56], too. They set up the auxiliary supervision of CNN to obtain the noise pattern and showed how different spoof mediums exhibited different noise patterns. In particular, the noise patterns of the live and fake images were different. End-to-end training of a CNN distinguished accurately between live and spoof accurately. Authors of [112] proposed a novel model, Spatio-Temporal Anti-Spoofing Network (STASN) to differentiate between live and spoofed faces. For anti-spoofing both temporal and spatial cues were used. The model used a new data synthesis method which provided a huge amount of training data. STASN combined with extensive training data provided improved performance when compared with the state-of-the-art methods.

2.6.5 Multi-modal methods

In an FR system, PAs occur in the visible light range. However, more cues on attacks are available from an other spectral image [113]. Multi-spectral face PAD approaches in recent literature are listed in Table. 2.4.

Jiang et al. [114] proposed a multi-spectral presentation attack detection approach to detect 3D mask and print attacks based on Visible Spectrum (VIS) and Near Infra-Red (NIR) images. Similarly, George et al. [115] used multi-channels (VIS, NIR and Thermal) and transfer learning to enhance performance. The method failed in identifying scenarios like prescribed glasses and facial hair attacks. The enhanced performance provided by extended-range imaging was utilized to detect PAIs in [116].

Table 2.4: Multi spectral anti-spoofing methods

Method	Modality	Attacks	Databases
Multi Level Image Fusing [114]	RGB, NIR	Print, 3D	CGIT PMT
Multi Channel CNN [115]	RGB, NIR, Thermal, Depth	2D, 3D	WMCA
Attention based Two Stream CNN[117]	RGB, MSR	Print, Replay	CASIA FASD, REPAY ATTACK, OULU
Multi Spectral Disguise Detection [24]	RGB, Thermal	Disguise	BVSD, IHTD
Multi Spectral Deep Embedding[116]	RGB, NIR, Thermal	Silicon Mask	XCSMAD
NIR Silent Liveness Detection Network Architecture [118]	NIR	Photo	Proprietary Dataset
Multi-modal FPAD with Spatial and Channel Attention [119]	RGB,IR,Depth	Photo	CASIA-SURF
Multiple Categories Image Translation GAN [120]	RGB, NIR	Photo, Video	CASIA-MFSD, REPLAY-ATTACK, Proprietary Dataset
Multi-Task cascaded CNN [121]	RGB, IR	Photo, Video	CASIA-MFSD, REPLAY-ATTACK, NenuLD

Kotwal et al. [116] addressed custom silicone mask-based impersonation PAD by deploying multi-channel inputs and CNN. Extracted feature vectors from CNN were classified using a logistic regression classifier. A two-stream convolutional neural network approach was set up in [117]. Two imaging spaces, RGB and Multi-Scale Retinex (MSR) were used in this approach to extract textural features and high-frequency information. The model was found to be insusceptible to illumination changes. A multi-spectral method to identify disguise was described by Dhamecha et al [24]. This method classified facial portions into patches of biometric and non-biometric based on the presence of disguise tools and then performed a recognition task.

Fan et al. proposed and evaluated NIR and VIS methods with NIR and VIS datasets respectively [118]. Through the experiments conducted, the authors verified the capability of NIR methods compared to VIS method. As per their observation, NIR provided more distinct features and NIR camera itself has some resistance to spoofing as it could not take images of replay attack using mobile and high-colour photos. Jiang et al. [120] utilised the cues from visible spectrum (VIS) and Near Infra Red (NIR) images. In this work, NIR images were synthesised using GANs [122] through image translation technique. The VIS and NIR pair gave cues for better spoof detection. Image translation using GAN provided required NIR image.

Liu et al. [121] proposed a PAD approach using IR and RGB images. The authors used a Multi-Task cascaded CNN (MTCNN). This approach exhibited lower responding time, making it suitable for real-world applications. Wang et al. [119] presented another multi-modal technique to detect spoofed faces. Using RGB, IR and depth modalities, the authors used an attention mechanism to capture information to detect spoofing. These three modalities and their combination trained a ResNet-18 model and were classified using the combination of softmax loss and center loss.

Authors of [77] used domain adaptation to transfer source domain information from VIS domain to the multi-spectral target domain. These multi-spectral methods were able to enhance spoof detection using reflection invariant cues obtained through extended imagery. However, these methods required an additional sensor along with VIS camera. Similar to existing other CNN based PAD methods, multi-spectral methods also required larger datasets with more attack variants in all modalities.

2.7 Hybrid methods

Local binary pattern (LBP) and its variants have been extensively used in handcrafted feature methods for FPAD. The authors of [123] proposed a novel fusion model to reduce training parameters by using the similarities between CNNs and LBP extraction.

This fusion network reduced the number of network parameters by using a statistical histogram. Nevertheless, the model failed to detect some specific types of attacks. Chen et al. [72] fused colour texture features with deep features from the images. Colour texture features were extracted by using rotation rotation-invariant local binary pattern (RI-LBP). These location features were fused with the global features extracted by using a ResNet model for classification. An SVM, with RBF kernel, classified these fused features to detect whether the face was authentic or spoofed. The experiments considered YCbCr and HSV colour spaces. In a comparison of grayscale, RGB, YCbCr, and HSV, the texture features, combined with the YCbCr and HSV colour spaces provided better detection results. The authors also presented a cross-dataset evaluation to show the generalisation capability of the method. The authors of [124] combined different handcrafted features including LBP, GDP, GLTP, LDIP, LGBPHS, and LPQ. These extracted features were classified by using the K-NN classifier. However, the model exhibited very low real-face detection accuracy regardless of the high (98.39%) fake-face detection accuracy. Moreover, this method only combined handcrafted features. Deep global features were not considered in this model.

Liu et al. [125] adopted a multi-modal data fusion strategy to identify fake faces. The model combined both low-level and high-level features from RGB and IR images for FPAD. This model exhibited generalisation against different conditions such as dim light, realistic face camouflage, static or motion pattern, etc. Because a nonlinear fusion method was used with multi-modal data, the generalisation was enhanced to some extent. The authors of [126] followed a dual cue fusion method to mitigate the error in FPAD. The framework had two streams. The first stream used facial images with background and the second stream used face images after facial area only. Fast Fourier Transform (FFT) was extracted from the facial images with background and these images were used to train the CNN model for FPAD. Simultaneously, the second stream carried out a colour space (HSV) transformation on facial area RGB images. Texture features were extracted from these images as input to SVM classifiers for FPAD. The decisions from both streams were combined to identify the PAs.

Younis and Abuhammad [127] proposed a hybrid fusion framework to address PAs by using multiple biometric modalities. The authors combined transfer learning and hand-crafted feature methods by using discriminant correlation analysis (DCA) and canonical correlation analysis (CCA). On the images, contrast adjustment was carried out to control intensity distribution. Histogram of gradient (HoG) features were extracted from these images. A multi-level fusion strategy was followed to incorporate multiple biometric modalities. In a dual stream fusion model, Fang et al. [128] used frequency domain features and complementary RGB features. This model included a hierarchical attention module as well as a multi-stage fusion strategy. A special attention module at the lower layers of CNN enabled the extraction of texture features. Similarly, a channel attention module at the higher layers extracted deep semantic features. Because both of these features are essential for better detection of attacks, this fusion model addressed generalisation through the decomposition of multi-level frequency.

Unlike other fusion models existing in the literature, the authors of [129] combined deep learning with serial fusion, as parallel fusion models have a longer response time. These multiple biometric modalities-based methods used Siamese neural networks. Deep networks were used for deep feature extraction and match score generation. Daniel and Anitha [130] proposed a new FPAD method, combining texture and image quality features. The image colour space was changed to HSV. Entropy-based colour texture features and image quality features were extracted from these HSV images. Later, these extracted features were concatenated and then classified. Even though this model combined different handcrafted features, it did not consider deep feature extraction to address PAs.

Xu et al. [131] used two lightweight networks to learn motion and texture cues in order to improve PAD. An element-wise weighing fusion strategy was followed in this model. In [132], the authors used camera-invariant feature learning while focusing on generalisation in FPAD. This framework learned both high-frequency and low-frequency information. A module in the framework carried out high-frequency domain camera-invariant feature decomposition. Another module in the framework performed image re-composition of both high- and low-level information. Classification results of both modules were fused together by using a weighting strategy to perform the final classification. Sharifi [15] proposed a decision-level fusion strategy to address FPAD. The author carried out feature extraction with a Log-Gabor filter. By using a nearest neighbours classifier, the scores were classified. Simultaneously, feature extraction and classification were performed by using a CNN model too. By using the OR rule, the decisions from the two modules were fused to get a final decision on the genuineness of the facial image.

Cai et al. [133] used meta-pattern learning, instead of hand-crafted feature extraction, to create a hybrid model to address FPAD. By using the hierarchical fusion module (HFM), RGB image and meta patterns were combined and passed to a CNN for further classification. Song et al. [134] proposed FPAD by using least squares weight fusion (LSWF) of channel-based feature classifiers. The authors utilised colour, texture, spatial domain, and frequency domain features extracted from different channel spaces along with convolutional features in this fusion method. To assign the optimal weights of classification score fusion, a least square weight fusion strategy was used.

Anand and Viswakarma [135] proposed a fusion method combining deep features and colour texture features. Extracted features were classified by using SVM separately. The probabilities from each model were fused to get the final probability. The authors of [136] utilised dynamic texture features and shape cues in a fusion method, to address 3D attacks. Geometric information used in this method was either extracted by the depth sensors or reconstructed from the RGB images. It also made use of a multi-modal dynamic fusion network and 3D model-guided data augmentation. This data augmentation facilitated data in different poses, which in turn assisted in training the network fully.

2.8 The Latest Methodological Breakthroughs in FPAD

It is essential to remain at the forefront of methodological advancements in the field of FPAD, which is constantly evolving. In the realm of FPAD, there have recently been groundbreaking developments that have the potential to revolutionize how generalization can be handled while addressing issues such as computational cost, privacy and legal issues, feature extraction, and fusion. With an aim to reduce computational cost while improving generalisation, the authors of [137], used a subset of frames from the dataset, with uniform sampling. It did not have to estimate the motion between adjacent frames as all frames are not necessary to detect facial movements for FPAD. Authors of [138] carried out a computational analysis of different CNN-based models. Fast RCNN was found to be effective with the best accuracy and efficiency in the PA detection task in this evaluation. However, the analysis was conducted using only two public datasets NUAA and CASIA which were limited in size and variations. Using datasets with a larger size and more variations can impact the analysis results differently.

Synthetic data generation [139] has been explored to address FPAD in recent research. In order to achieve domain generalisation, synthetic data was used to train deep ensemble models [140]. In this video-based data augmentation method, the authors considered both spatial and temporal features. As the model was able to learn spatiotemporal features from video, it can also overcome the inability of images to provide more cues for PA detection. Authors of [141] created a new T-shirt Face Presentation Attack (TFPA) dataset, using 8 subjects and 100 synthetically generated facial images. This multi-modal dataset with RGB and depth images has 1608 samples. Among the SOTA methods evaluated, the anomaly detection method showed better performance. It also showed that T-shirt attacks could be used against FR if the actual face was covered. However, the cross-dataset generalisability was not evaluated with this dataset.

Another recent trend in addressing FPAD generalisation is collaborated learning methods including federated learning. Such methods can overcome legal and privacy issues related to sharing multiple source domains among different entities in practical scenarios. FedSIS [142] used ViT architecture. It used a feature augmentation strategy by utilising intermediate features of ViT. A well-generalised cross-domain performance was achieved using its method while preserving data privacy. FASS [143] used image quality features and deep features to form a late fusion model. No-reference image quality features were classified with SVM and RF, and the classification results were combined with ResNet-50 classifier results. This score-level fusion exhibited PA detection with generalisation comparable to existing SOTA methods.

Authors of [144] addressed the FPAD problem using Multi-Scale Colour Inversion (MSCI) of images and early fusion of features. In this two-stream method, one stream was used to extract features from RGB images, while the other stream extracted features from MSCI images and then these features were combined to detect the PAs. Multi-scale colour Inversion provided face reflection features. A Paralleled convolutional block attention module (PCBAM) was used to make the network give more attention to regions which were relevant to the PA detection task. The model exhibited generalisation across different datasets showing the potential to detect unseen attacks. Geometric facial dynamics from dense landmark predictions were explored in [145] using Geometry-Aware Interaction Network (GAIN), to detect PAs. The authors also designed a cross-attention strategy in order to integrate the extracted geometric features with editing SOTA methods which in turn improved detection performance both in intra-dataset and cross-dataset evaluations showing improved generalisation.

Multi-modality in FPAD generally utilised extended imagery. In contrast to this, authors of [146], used sensors including a speaker and microphone other than the camera. Thus a method combining visual and auditory modalities was used to address PA detected. To facilitate better feature learning, a hierarchical cross-attention mechanism also was incorporated in this method. One breakthrough in multi-modal FPAD is the flexible-modal framework, which followed "train one for all" [147]. The model was trained using a combination of RGB, Depth and IR data. The trained FPAD model is tested using four modality combinations including RGB, RGB+Depth, RGB+IR, and RGB+Depth+IR. The model was proposed to mitigate the redundancy and inefficiency associated with multi-modal training scenarios. However, through evaluations, it was found that for flexible modal FPAD, the training set should have all the modalities simultaneously. Then only, the model can be evaluated using any of the modality combinations. In [148], the authors leveraged visual saliency to improve the performance of the PA detection. In each frame, saliency information is extracted based on the difference between the Laplacian and Wiener filter outputs. This approach highlighted the significance of enhancing the representativeness and diversity of a training dataset by prioritizing the inclusion of the most salient images or regions within the images. This strategy holds the potential to yield more efficacious models in the context of PA detection. Nonetheless, a limitation identified was that it might not be able to capture all fine details in longer videos.

Authors of [149] explored the possibility of vision-language pre-trained models in improving FPAD generalisation. As part of their research, the authors also investigated whether self-supervision techniques could be incorporated into the adaptation of VLP models for FAS in order to overcome the large domain gap and limited availability of training data in FAS. Using this method, Generalisation was further enhanced by aligning image representations with text representations produced by the text encoder. Advanced Multi-Perspective Feature Learning network (AMPFL) [150] utilised discriminative features from multiple perspectives to improve PA detection task. This approach used a Multi-Perspective Feature Learning module to capture cues specific to various perspectives, whereas a Multi-Perspective Feature Fusion module combined perspective-specific cues with universal cues, which facilitated FAS with a more comprehensive set of information.

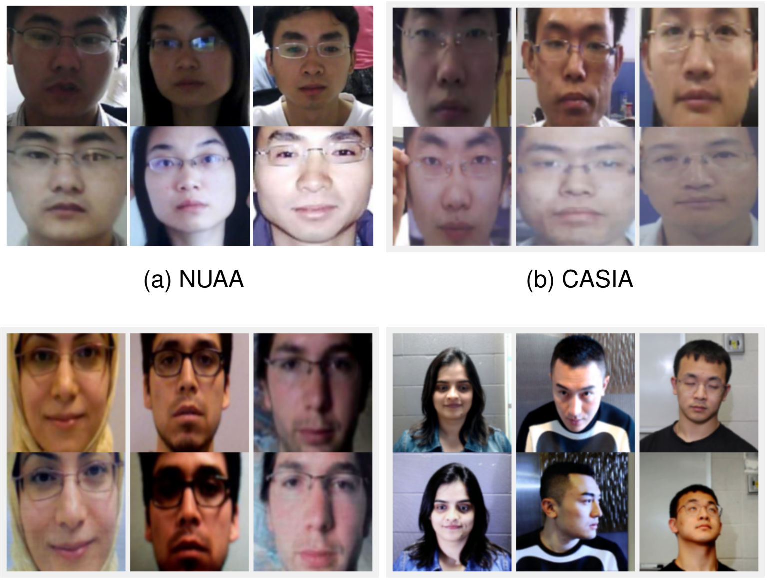
2.9 Datasets

Datasets have a pivotal role in the performance of any presentation attack detection method. The generalisation of PAD relies on variance in samples of a dataset. Access to a wider variety of PAs facilitates the learning of more attack features during the training process. This eventually leads the system to detect the PAs of a wide range. Samples of print attack images with different illumination conditions from NUAA imposter are shown in Figure. 2.13. For each dataset, the upper row shows the real facial images and the corresponding presentation attack images are in the lower row. (a) NUAA imposter dataset [10], (b) CASIA-FASD [41], (c) Replay Attack [46], and (d) SiW[102]. Other attack variants include replay attacks, 3D mask attacks and their variants. Both print and replay attacks from CASIA, Replay Attack and SiW are also shown in Figure. 2.13.

It is evident from Table. 2.5 that existing datasets consist of more 2D attacks than

3D attacks [151, 5]. However, diverse novel attacks are increasing with progressive technology. Dataset diversity is decided by PAs and their variants. Factors such as environment, recording set up, illumination, pose, expression and spoofing medium also affect the dataset content.

The deep learning model is a neural network model with several layers and parameters between output and input. Figure. 2.3 illustrates a simple neural network model. As a general rule, neural networks are composed of three layers: the input layer, the hidden layer, and the output layer. As deep learning models use neural networks with more hidden layers, they are commonly referred to as deep neural networks. Deep refers to the fact that there are more hidden layers between the input and output layers. An ordinary neural network model may contain two or three hidden layers, whereas a deep neural network model or a deep learning model may contain 100 or more hidden layers.


(a) NUAA

(b) CASIA

(c) Replay Attack

(d) SiW

Figure 2.13: Real and fake facial images from four public datasets.

Dhamecha et al. [24] developed a multi-spectral dataset for disguise attacks called IIITD: In and Beyond Visible Spectrum Disguise (I^2BVSD). This dataset consists of 75 subjects with various disguise accessories. Both visible and thermal spectra were considered for data acquisition. The authors introduced distinct disguise variants for the dataset as:

- Without disguise

- Variations in hairstyles
- Variations due to beard and moustache
- Variations due to glasses
- Variations due to cap and hat
- Variations due to mask
- Multiple variations

Table 2.5: Face Anti-Spoof (FAS) datasets

Dataset	**Year**	**Subjects**	**Samples**	**Modality**	**Attacks**
NUAA [10]	2010	15	12,641	RGB	Print
CASIA-MFSD [41]	2012	50	600	RGB	Print, Replay
Replay-Attack [46]	2012	50	1,200	RGB	Print, Replay
YMU [152]	2012	151	604	RGB	Makeup
ERPA [153]	2013	5	86	RGB, Depth, IR, Thermal	3D Silicon/resin Mask
MIW [154]	2013	125	154	RGB	Makeup
MLFP [155]	2013	10	1,350	RGB, Thermal	3D Latex, Paper Mask
GUC-LiFFAD [156]	2015	80	4,826	RGB	Print, Replay
MSU-MFSD [43]	2015	35	440	RGB	Print, Replay
3DFS-DB [157]	2016	26	520	RGB, IR	2D/3D Mask
3DMAD [158]	2016	17	255	RGB, Depth	3D Mask
HKBU MARs [159]	2016	12	1,008	RGB	3D Rigid Mask
MSSPOOF [160]	2016	21	4,704	RGB, IR	Print
Replay-Mobile [161]	2016	40	1,030	RGB	Print, Replay
BRSU [162]	2017	50	141	RGB, IR	3D Masks, Facial disguise
CIGIT-PPM [114]	2017	72	93,358	RGB, IR	Print, 3D Mask
EMSPAD [163]	2017	50	14,000	7-band multi-spectral data	Print
MIFS [164]	2017	107	416	RGB	Makeup
Oulu-NPU [165]	2017	55	5940	RGB	Print, Replay
SMAD [166]	2017	From internet	130	RGB	3D Silicon Mask
CS- MAD [45]	2018	14	308	RGB, IR, Depth, LWIR	3D Silicon Mask
DFW [167]	2018	1000	11,155	RGB	Disguise
Rose-Youtu [168]	2018	20	3350	RGB	2D, 3D
SiW [102]	2018	165	4620	RGB	Print, Replay
WMCA [38]	2018	72	6716	RGB, Dept, IR, Thermal	Print, Replay, 2D/3D Mask
3DMA [169]	2019	115	920	RGB, IR	3D Mask
AIM [170]	2019	72	456	RGB	Makeup
CASIA-SURF [171]	2019	1000	21000	RGB, Depth, IR	Print, Cut
I^{2}BVSD [24]	2019	75	681	RGB, Thermal	3D Facial Disguise
LCC FASD [172]	2019	243	18827	RGB	Photo
PR-FSAD [173]	2019	30	127440	RGB	Print, Replay
SiW-M [101]	2019	493	1,630	RGB	Print, Replay, 3D Mask, Makeup
WFFD [174]	2019	745	2300	RGB	Wax figures
CASIA SURF CeFA [175]	2020	1,607	23538	RGB, Depth, IR	2D, 3D
CelebA-Spoof [176]	2020	10177	625537	RGB	Print, Replay, 3D, Paper cut
TFPA [141]	2023	108	1608	RGB, Depth	2D

Disguised Faces in the Wild (DFW) dataset [177] is a similar dataset with disguised face attacks. It has images of 1000 subjects. A total of 11,155 face images of real-world disguise variants obtained from internet sources, formed this dataset. Bhattacharjee et al. [45] created a new Customised Silicon Mask Attack Dataset (CS-MAD) and verified the vulnerability of face biometric system using the dataset. The boost in technology made the manufacturing process of mask easier and cheaper and a number of recent datasets incorporate 3D attacks. Different mask attack datasets are described in Table. 2.7.

Zhang et al. [171] developed a new dataset, CASIA-SURF which was larger than existing datasets in size. The dataset consists of three modalities which are VIS, IR and depth. It has 21,000 sample videos from 1000 subjects. The authors of [170] formed a novel Age Induced Makeup (AIM) dataset. 456 samples using age progressive makeup type from 75 subjects were considered while forming the dataset. Liu et al. [178] formed a Spoof in the Wild (SiW) dataset introducing more spoofing medium and recording settings with photos of 165 subjects. The authors of [169] developed a dataset for 3D Mask Attacks (3DMA) based on VIS and NIR. Xiao et al. developed this dataset in order to apply more variance in lighting distance and illumination deploying various methods. 920 videos of 67 subjects were included in the dataset. There were 48 3D mask variants used to create this dataset.

Table 2.6: Multi-spectral datasets

Database	Year	Modality	Samples	Attacks
ERPA [153]	2013	RGB, Depth, IR, Thermal	86	3D Silicon/ resin Mask
MLFP [155]	2013	RGB, Thermal	1350	3D Latex/ Paper Mask
I^{2}BVSD [179]	2014	RGB, Thermal	681	3D Facial Disguise
3DMAD [158]	2016	RGB, Depth	255	3D Mask
MSSPOOF [160]	2016	RGB, NIR	4704	Print
EMSPAD [163]	2017	7-band multi-spectral data	14,000	Print
BRSU [162]	2017	RGB, 4 SWIR bands	-	3D Masks, Facial disguise
CIGIT-PPM [114]	2017	RGB, NIR	93358	Print, 3D Mask
WMCA [38]	2018	RGB, Depth, IR, Thermal	6716	Print, Replay, 2D/ 3D Mask
3DMA [169]	2019	RGB, NIR	920	3D Mask
CASIA-SURF [171]	2019	RGB, Depth, IR	21000	Print, Eye-Cut photo
CASIA-SURF CeFA [175]	2020	RGB, Depth, IR	23538	2D, 3D
HQ-WMCA [180]	2020	RGB,Depth,NIR,SWIR,Thermal	58080	Print, Replay, Partial, Mask
PADISI-Face [181]	2021	RGB, Depth, NIR, SWIR, Thermal	121740	Print,Rreplay, Mask, makeup/tatoo, partial
LDFAS [182]	2022	RGB, LiDAR	2880	Print, Replay, Mask
UVLD [183]	2023	RGB, Ultrasound	36258	Replay
Echoface-Spoof [146]	2023	RGB, Acoustic	249352	Print, Replay
TFPA [141]	2023	RGB, Depth	1608	2D

Table 2.7: 3D mask datasets

Database	Year	Subject	Sample	Material
3DMAD [158]	2013	17	255	Paper, hard resin
3DFS-DB [157]	2016	26	520	Plastic
HKBU-MARs [159]	2016	12	1008	Rigid (Two different manufactures)
BRSU [162]	2016	137	141	Silicon, plastic, resin, latex
SMAD [166]	2017	From internet	130	Silicon
MLFP [155]	2017	10	1350	Latex, paper
ERPA [153]	2017	5	86	Resin, silicone
WMCA [38]	2019	72	1679	Rigid, silicone, paper
WFFD [174]	2019	745	2300	Wax figure
CIGIT-PPM [114]	2019	72	93358	Leather, rubber, plastic
3DMA [169]	2019	115	920	48 Variations of masks
SuHiFiMask [184]	2023	101	10,195	resin, silicone, plaster, headgear, head moulds

Emphasizing on video replay attack, Timoshenko et al. created a larger dataset. The Large Crowd Collected Facial Anti-Spoofing Database (LCC FASD) in [172] has more variance in devices deployed for recording and replay. The dataset has 1942 real faces and 16885 attack samples. In [173], the authors introduced a novel dataset Pattern Recognition Face Spoofing Advancement Dataset (PR-FSAD) for spoof detection which emphasizes on variations in angle and distance. 42,480 real and 84,960 fake samples from 30 subjects used to construct the dataset. A new dataset, Digital Forensic - Face Presentation Attack Detection (DF-FPAD) was created for the evaluation process of a presentation attack detection framework using this textural noise in [185]. The dataset was made using higher-quality images of fake and genuine faces under controlled conditions.

2.10 Evaluation Metrics

Face PAD is commonly considered as a binary classification problem. Various performance-associated metrics are used to evaluate the performance. Chingovska et al. detailed about measuring face PAD as a binary classification problem [186]. Since these binary classification systems are provided with two classes of input, they are normally termed positive and negative classes. Their performance is evaluated by the types of errors committed and the method to measure them. False Positive and False Negatives are the errors exhibited by binary classification systems. Normally recorded error rates are False Positive Rate (FPR) and False Negative Rate (FNR). FPR is the ratio of FP to the total number of negative samples and FNR is the ratio of FN to the total number of positive samples.

In biometric verification systems, the performance relies upon the acceptance or rejection of the sample. So the terms False Positive Rate (FPR) and False Negative Rate (FNR) are replaced by False Acceptance Rate (FAR) and False Rejection Rate (FRR), respectively [187]. As there is the matching process involved in the verification task, FAR and FRR are often described as False Match Rate (FMR) and False Non-Match Rate (FNMR) [188]. Anti-spoofing systems function on the concept of acceptance and rejection. So usually PAD systems use FRR and FAR. The ratio of incorrectly accepted spoofing attacks defines FAR, whereas FRR stands for the ratio of incorrectly rejected real accesses [186].

Presentation Attack Detection (PAD) follows ISO/IEC DIS 30107-3:2017 [189] to evaluate the performance of the PAD systems [190]. Authors of [5] described evaluation metrics used for testing different scenarios in a PAD system. The most commonly used metric in anti-spoofing scenarios is Half Total Error Rate (HTER) [186]. HTER is found out by calculating the average of FRR (ratio of incorrectly rejected genuine score) and FAR (ratio of incorrectly accepted zero-effort impostor). FAR is associated with SFAR (ratio of incorrectly accepted spoof attacks). PAD methods used Equal Error Rate (EER) to test reliability [5]. EER is a specific value of HTER at which FAR and FRR have equal values.

Table 2.8: Commonly used evaluation metrics in face PAD

Metrics		**Equation**
False Acceptance Rate	FAR	$\frac{FP}{Fake\ samples}$
False Rejection Rate	FRR	$\frac{FN}{Genuine\ samples}$
Equal Error Rate	EER	$(FRR = FAR)$
Half Total Error Rate	HTER	$\frac{FAR+FRR}{2}$
Attack Presentation Classification Error Rate	APCER	$\frac{FP}{FP+TN}$
Bona fide Presentation Classification Error Rate	BPCER	$\frac{FN}{FN+TP}$
Average Classification Error Rate	ACER	$\frac{APCER+BPCER}{2}$

While evaluating some methods, metrics mentioned as per ISO standard in [189] were used. They were Attack Presentation Classification Error Rate (APCER), Normal Presentation Classification Error Rate (NPCER) and Average Classification Error Rate (ACER). NPCER is identical to the Bona fide Presentation Classification Error Rate (BPCER). A Face PAD is evaluated in terms of classification of attacks and real face, intra-dataset performance and cross-dataset performance [56]. BPCER and APCER measure bona fide and attack classification error rates respectively. ACER evaluates the intra-dataset performance, whereas HTER scales cross-dataset performance [189]. Commonly used metrics [191, 39, 5] in face anti-spoofing are listed in Table. 2.8. Computational latency stands as another significant parameter [192] that can be employed for comparing the state-of-the-art method. Latency holds a crucial role in the real-time deployment of models.

2.11 Challenges and future directions

Despite the recent progress in presentation attack detection methods, unseen attack detection is still considered a challenging problem. Existing methods showed promising results when evaluated using a specific type of attack under a controlled environment or using public datasets. PAD models trained used predefined attacks also show promising results, however, such models tend to be biased toward these types of attacks [193]. While machine learning models perform well on samples taken from within the same distribution as the training set, that performance is not maintained across different datasets or in new conditions. In other words, generalising performance across a wide range of attacks and across different datasets is still considered an inherently challenging problem. This can be partly attributed to common computer vision challenges such as the distance of the subject to the camera, image resolution, light [83], pose variations and others. This suggests strongly that, PAD in an uncontrolled environment requires further research efforts [194, 195].

One of the key challenges to progress research and development of PAD methods is the large number of ways that such attacks can be performed. It remains impractical to compile a dataset that captures all current attack variations regardless of their type (e.g. 2D, 3D attacks). It is impossible to predict the varieties of attacks that new technological advances will bring in the future. The literature shows that compared to existing 2D attack datasets, 3D attack datasets and multi-spectral datasets are scarce with fewer subjects to compare to image classification and face recognition datasets. More datasets in the public domain are required to progress research in this area. In particular, datasets that capture novel attacks using recording devices, and other new emerging technologies [171].

The inclusion of temporal features, such as motion or rPPG, for auxiliary supervision is another challenging task in face PAD. The majority of auxiliary methods in face PAD used spatial features, especially depth as an auxiliary feature. These have considered a single frame for detection. Limited research has been conducted to utilise temporal features for auxiliary supervision. This may be partly attributed to computer processing requirements and the need for rapid processing in face recognition systems. Multiple frames with longer duration have to be processed to deploy temporal features for auxiliary supervision. Hence, multiple frame-based models increase processing time within the face recognition systems [106]. As technology advances, however, temporal features might increase accuracy in PAD, and this research area should not be neglected.

Remodelling face presentation attack detection as one class classification approach has provided impressive results in unseen attack detection. Hence, this approach is a promising future research direction. Delving further into anomaly detection, few-shot learning, zero-shot learning, and domain generalisation is recommended for enhancing unseen attack detection. Combining this with further investigation into auxiliary supervision with more spatial and temporal features would provide a powerful, new research direction. Recent research has investigated multi-spectral data augmentation using image translation and GANs. This has provided new methods which utilize multi-spectral cues without the need for physical auxiliary sensors. GANs have also been used for learning generalised features over multiple domains in feature space. Hence, further study with GANs in anti-spoofing might provide some way of generalising presentation attack detection over unseen attacks.

2.12 Conclusion

Presentation attacks continue to pose a challenge for the research community despite recent and significant progress in the development of detection methods. Methods such as anomaly detection, domain generalisation, few-shot learning, zero-shot learning and others have shown some promising results. In this chapter, a comprehensive review of existing presentation attack methods is presented, along with an assessment of challenges and possible research directions. FPAD presents a number of challenges, including the absence of a common dataset protocol, the need to include a wide range of datasets, and the need to extract specific features relevant to FPAD. Based on these findings, FPAD needs to be explored using various public datasets in order to determine the impact of different dataset protocols, increase the variance of the training dataset, and extract FPAD-specific features on generalisation. Rather than official dataset partitions, custom data partitions have been used as a way to increase the variance of training sets. The upcoming chapter discusses the impact of this practice on the FPAD.

Chapter 3

Custom Dataset Partitions: An Impact Analysis on FPAD

In the previous chapter, we saw how deep learning FPAD methods usually outperform traditional machine learning methods. There is a discrepancy, however, in terms of how the datasets are partitioned. This chapter presents an experimental framework to investigate how different train-test partitions and variance in training data affect model performance with the NUAA dataset [10] for PA detection. The results show that using different partitions of this dataset results in different models with different performances. The main findings in this chapter appeared in the proceedings of the 22nd International Conference on Engineering Applications of Neural Networks (EANN 2021) [196].

3.1 Overview

Face presentation attack detection (FPAD) uses either manual feature extraction or deep learning [5]. These features include texture, image quality, temporal cues, and life signs [197]. For the detection of PAs, the manually extracted features have been classified using machine learning classifiers such as Support Vector Machines (SVMs) and Random Forests (RFs) [198]. The inherent feature learning capacity of deep neural networks has recently been utilised in several different approaches for the detection of PAs. [1]. Hence, manual feature extraction methods have been replaced with deep learning techniques, providing better performance in PA detection. During the intra-dataset evaluation, the majority of these methods were found to be effective. In cross-dataset evaluations, however, the performance of these models has been shown to deteriorate substantially, indicating a low degree of generalisation capability. Consequently, some models may not perform as expected in real-life situations.

Like any other image classification problem, the dataset plays a critical role in FPAD generalisation. In order to achieve better generalisation a dataset should contain more data samples and exhibit high variance [14]. Different illumination, setting, recording devices, ethnicity, spoofing medium, and materials all contribute to PAD dataset variance. Nevertheless, some existing Face Anti-Spoof (FAS) datasets have limited variance and size. These limitations also create a biasing of the model towards the training set features, which affects the model's generalisation capability.

NUAA imposter dataset [10] is a publicly available, FAS dataset of manageable size. This dataset has a larger test set than the training set, unlike usual datasets. The train and test sets were recorded on different days. Hence, there is no overlapping between test and train images from different sessions, and the original dataset test/train partitions were appropriately disjoint. Also, the images in this dataset have various illumination conditions. This session-wise recording with varying light conditions may have made this dataset suitable for evaluating the generalisation capability of a model. Recent practice, however, does not use the original disjoint test/train partition, instead using customised partitions to give a larger training set with smaller test and validation sets. Hence, based on the above factors, an experimental framework was created to detect photo attacks using deep learning and extracted features. Two different dataset partitions were tested with each model.

3.2 Customising NUAA dataset partitions for FPAD

In attaining generalisation in FPAD, the dataset plays a crucial part. Dataset variance in terms of spoofing medium, illumination, ethnicity, settings, recording devices and materials along with a number of samples contribute to this improved generalisation. NUAA Imposter dataset [10] is one such dataset which has varying illumination conditions. This dataset had been extensively used in FPAD model evaluation. The NUAA dataset has a proper disjoint train-test partition in terms of recording sessions as they were captured on three different days. Hence, NUAA is capable of testing the unseen attack detection performance of a model. Nevertheless, in recent research, instead of using train-test partition with disjoint distribution (3491 train and 9123 test images) as in [10], the majority of models use varying train-test partitions to train and evaluate models. While the actual partition has a smaller train set than the test set, the majority of the customised partitions have larger training sets with images from all sessions. But in real partition, the train set images are only from session 1 and session 2, whereas test images are from session 3.

Many of earlier FPAD models using manual feature extraction, used NUAA with the actual partition as in [10]. The models such as Multi-scale LBP [47], a combination of Karhunen-Loeve Transform (KLT), 2D Fourier Transform, DoG with LBPV and MLBP combined with image quality features such as colour moment, R-G deviated texture utilised this train-test partition. In intra-dataset evaluation, these models even performed well in spoof detection. However, lately, FPAD models were evaluated using different partitions on NUAA (Table. 3.1).

Table 3.1: Customised NUAA train-test partitions used in recent research

Author	**Method**	**Accuracy**	**Train-Test Ratio**
Matta et al. [47]	Multi-scale LBP+SVM	98	3491:9123
Li et al. [65]	KLT+CLBP+2D FT+SVM	95.21	3491:9123
Hasn et al. [9]	DoG+LBPV+SVM	99.22	3491:9123
Song and Ma [134]	LTP + LBP + R-G extractor+ COLOR MMT	98.49	3491:9123
Fahn et al. [197]	IDA+LBP+RF	99.04	90:10
Satapathy & Livingston [199]	Xceptio-Inception/Reduction CNN	100	80:10:10 & 60,20,20
Parveen et al [200]	DLTP	94.5	1745:1746:9123
Luan et al. [201]	Recaptutred Feature Extraction	98.8	50:50
Alotaibi et al. [57]	Non-linear Diffusion+CNN	99	3491:9123

Parveen et al. [200] proposed a manual feature extraction method, Dynamic Local Ternary Pattern (DLTP) to identify fake faces. The authors proved that their model performance was comparable with state of the art method. However, the NUAA dataset partition used in their model evaluation and compared models were different. The authors partition NUAA actual train set to train and validation set with 1745 and 1746 images but retained the test set as it was in the actual test set. Yilmaz et al. [54], compared various dimension reduction methods in the FPAD context. In their experiments, the authors used 5-fold cross-validation on the NUAA dataset. Simulated Annealing (SA) was found to be better in performance, through their experiments.

Fahn et al. [197] evaluated both handcrafted feature methods and CNN-based methods using the NUAA dataset. The authors implemented LBP, image distortion analysis and CNN. For image distortion analysis, specular reflection, chromatic moments, colour diversity, and sharpness were considered. For classification, random forest classifiers and deep neural networks (DNN) were used. Experiments were carried out using 10-fold cross-validation on the NUAA dataset. Even though the model showed excellent intra-dataset performance with NUAA, their cross-dataset performance, with respect to NUAA was not at the desired level.

The authors of [202] combined texture analysis and CNN to address FPAD. Based on the additive operator partitioning (AOS) scheme and triagonal matrix block solver algorithm, non-linear diffusion was applied to the NUAA dataset images. These images were used to evaluate the proposed five-layer CNN (CNN-5), ResNet50, and Inception V4. Among the tested models, Inception V4 performed better than other models. In their experiments, the authors used actual NUAA partition unlike other CNN-based methods evaluated on NUAA.

The majority of the manual feature extraction methods used the NUAA dataset with the train-test partition as per dataset protocol [10]. However, in recent research, both handcrafted feature methods and CNN-based methods had different partitions on NUAA. The authors of [203] partitioned the NUAA train set into train and validation sets of equal size, while retaining the test set as per the actual partition. Luan et al. [201] combined the train test and test from the actual partition and divided the whole dataset in a 1:1 ratio to get the train and test set. These models proved comparable to those feature extraction models, which used actual NUAA train-test partition.

Authors of [199] presented a CNN FPAD model and compared the model performance using two different partitions on the NUAA dataset. The authors used 80:10:10, and 60:20:20 random combinations of the NUAA dataset for the train:test:validation partitions, which were different from dataset specifications. In the existing research, different models trained on different partitions of NUAA were compared to demonstrate their superiority among the state-of-the-art methods. Since the dataset variance substantially changes with different partitions, it is very crucial to match the dataset partition along with performance measurement in this case.

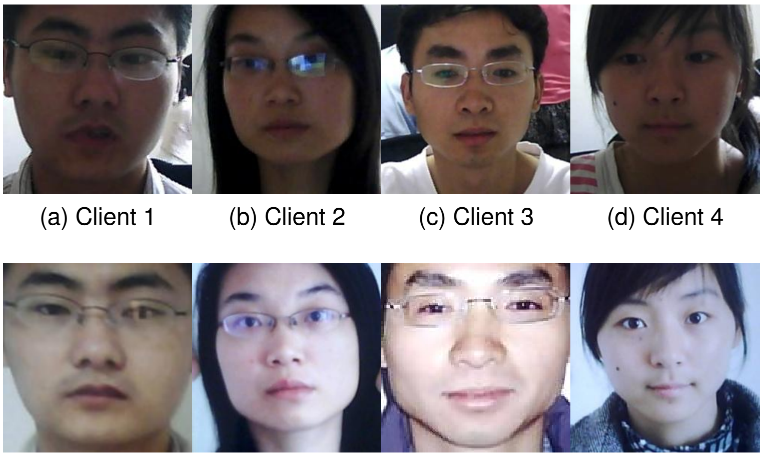
Authors of [57] presented a model using the diffusion technique. This technique reduced the processing time in the experiment, which used the AOS scheme for diffusion. The NUAA dataset showed the highest accuracy (99% with HTER=.098%). However, as the number of iterations and time step values were increased, the model performance of this dataset degraded. As the time step increases, features such as edges, location, fade-out, and iterations blur the image. These factors would influence the model performance. Beham and Roomi [204] adopted a depth map-based method to detect spoofed faces. The authors extracted Aggregated Local Weighted Gradient Orientation (ALWGO) features from the depth map. NUAA, CASIA, and Replay-Attack datasets were used for evaluation. The authors evaluated this model by changing the number of training images from 1 to 3148. Luan et al. [201] addressed the FPAD problem by analysing specular reflection ratio, hue channel distribution and blurriness. The model used an equal number of images for training and testing and had an accuracy of 98.80%, which was better than the other feature extraction methods considered for comparison.

In [198], the authors presented a multi-texture analysis method. The proposed method, Completed Local Binary Pattern (CLBP) and Completed Local Binary Pattern Normalized Histogram Fourier (CLBP_NHF), performed better than LBP and MLBP. A colour chromatic moment feature extraction method was adopted in [205] to distinguish between real and fake images in FR. RGB images were converted to HSV and YCrCb colour spaces. The mean and standard deviation of pixels of images for each channel in these colour spaces were calculated. These statistical features of the image were also considered as colour chromatic moment features. For this model, the authors used actual NUAA train-test partition as in [10].

Sengur et al. [206] also proposed an FPAD method using pre-trained CNN models and trained the model with the actual NUAA train-test partition. AlexNet and VGG16 models were used for feature extraction. The SVM classifier was used to classify the features extracted using fully connected layers ('fc6' and 'fc7') of the pre-trained models. These layers provided a feature vector with a dimension of 4096. Concatenating features from fully connected layers from both models provided better accuracy than individual feature vectors from each model's fully connected layers. However, the results were comparable with the performance of the state-of-the-art methods.

3.3 NUAA Imposter Dataset

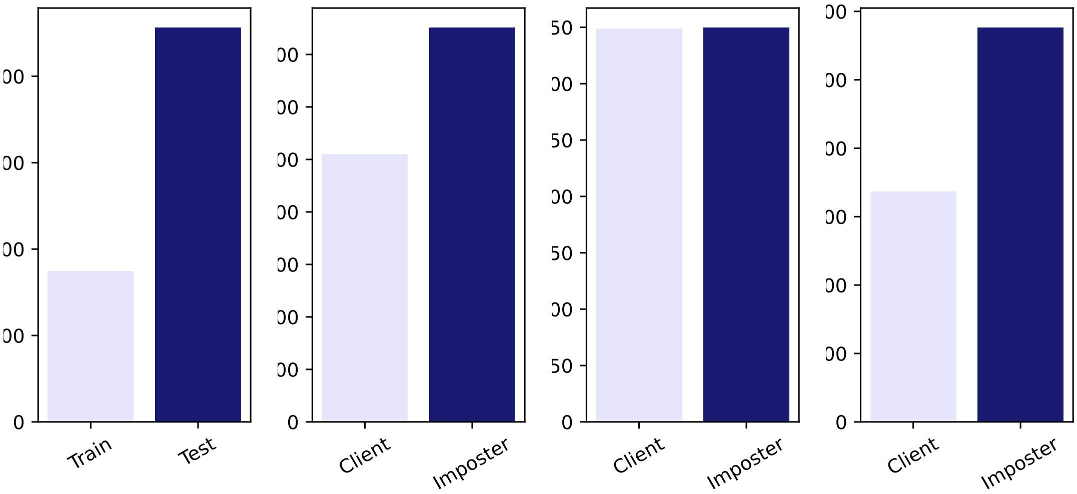
NUAA Imposter Database [10] has client and corresponding photo attack samples. Fifteen subjects were involved in constructing the dataset in three different sessions. The sessions differed in terms of lighting conditions, and time. The subjects were of different gender and age. Unusually the official test set is much bigger than the training set. The training set has 3491 images, and the testing set has 9123 images. The training set includes real and fake images of 9 subjects, whereas the test has images from 15 subjects. Fig. 3.1 shows real and fake image samples from the NUAA dataset. The dataset has raw images with a size of 640X480, face-detected images, and normalised face-detected images of 64X64 pixels. The raw images have the background information with face images. NUAA dataset images were recorded in three sessions. Session 1 and session 2 images were included in the training set, and test set images were taken in session 3. Fig.3.2 shows the official dataset partition, class distribution, class distribution for the train set and class distribution for the test set for the NUAA dataset.



(a) Client 1 (b) Client 2 (c) Client 3 (d) Client 4

(e) Imposter 1 (f) Imposter 2 (g) Imposter 3 (h) Imposter 4

Figure 3.1: Client and corresponding imposter image samples from NUAA dataset



(a) Dataset partition (b) Class distribution (c) Class distribution-train (d) Class distribution-test

Figure 3.2: Official dataset partition and class distribution in NUAA dataset.

Prior research split the NUAA dataset into train and test partitions according to the dataset protocol and utilised manual feature extraction methods to detect PAs. In subsequent research, however, both CNN-based and manual feature extraction-based methods used different partitions on NUAA. In [203], the authors divided the training set into two equal sets and used them as training and validation sets. They used a test set as per the dataset specification. The authors of [201] used a 1:1 train-test images ratio. Behma et al. [207, 204] carried out experiments with 5-fold cross-validation on the entire dataset.

The authors of [54] also carried out 5-cross validation on NUAA. Authors of [199] carried out extensive experimentation and evaluation on various models, which exhibited excellent performance with NUAA. However, the authors used 80:10:10, and 60:20:20 random combinations of the NUAA dataset for the train:test: validation partitions, which were different from dataset specifications. In the existing research, different models trained on different partitions of NUAA were compared to demonstrate their superiority among the state-of-the-art methods. Since the dataset variance substantially changes with different partitions, it is very crucial to match the dataset partition along with performance measurement in this case.

Since train and test sets in the NUAA dataset were recorded in different sessions, there is no session-wise overlap. However, the train-test partition used in [199, 207, 204] will introduce this overlapping. Illumination, devices, and recording settings of the dataset images affect the generalisation capability of the model. Even though models acquired high intra-dataset accuracy in such cases, the cross-dataset performance may deteriorate as in [197].

3.4 FPAD using different dataset partitions

An experimental framework was constructed to investigate the impact of dataset partition on pre-deep learning models and manual feature extraction methods in detecting PAs. Extracted features were then classified using SVM and RF classifiers as shown in Fig. 3.3. Two different train-test partitions on the NUAA dataset were considered in these experiments. They are actual NUAA partition with 3491 train images and 9123 test images [10] and customised partition with 80% images as a train set. In deep learning models, the customised test-train partition uses 10% validation and 10% test data, whereas, in manual feature extraction methods, test data was 20%. However, both scenarios had 80% train data, which was managed by combining actual train and test sets from NUAA and then partitioning it. The actual NUAA partition has disjointed train and test sets in terms of recording sessions. However, combining the entire dataset and customising the train and test partition caused the loss of non-overlapping characteristics of train and test sets.

Extracted features include Local Binary Pattern (LBP), colour texture analysis, colour moment, colour diversity, and variance. These features and their combinations were classified using SVM with linear kernel, SVM with radial basis function kernel (SVM-rbf), and RF. Among the manual feature extraction methods in existing literature, LBP and its variants have been used extensively in FPAD [5, 208]. Hence, this classical feature-engineering method was considered. However, LBP converted the image into gray-scale and did not utilise colour texture variations. In order to incorporate colour texture variations in fake and genuine images, colour texture analysis using colour LBP [8] was extracted and classified in here.

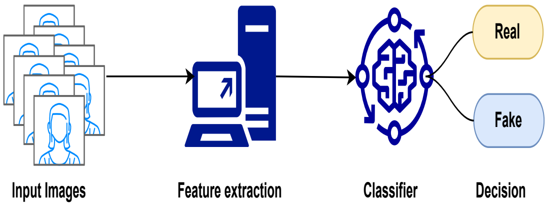


Figure 3.3: FPAD using handcrafted feature method

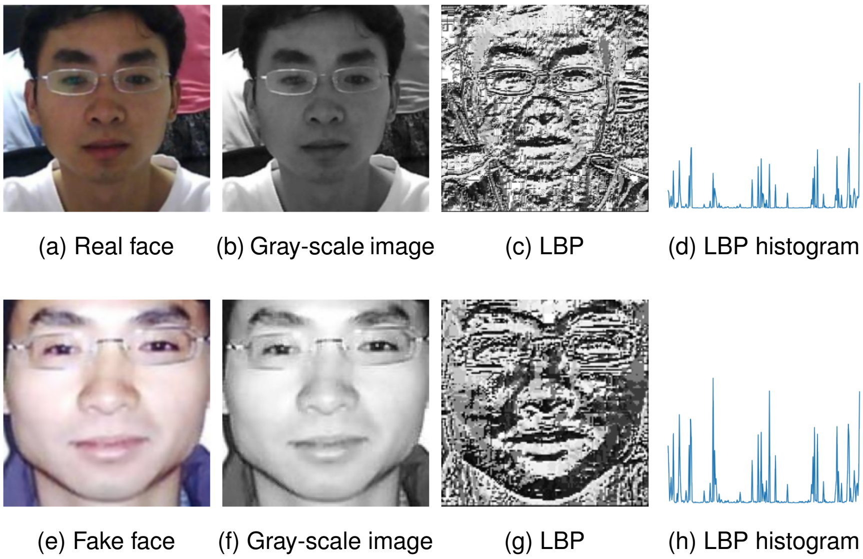
Apart from texture analysis, image quality factors also provide cues to detect spoofed faces. Hence, image quality factors such as colour moment, colour diversity and variance [197] were extracted in the experiment. The combination of these features with LBP was classified with SVM-linear, SVM-rbf and RF. Also, pixel values from images were extracted manually and classified using above mentioned three classifiers.

The aforementioned classical feature-engineering methods were compared with both shallow and deep CNN models for spoof detection. These models used two different train-test partitions to compare their performance in FPAD. The first partition has 3491 train images which were captured in session 1 and session 2, whereas in the other partition, the training set consists of 80% of the whole dataset. These images were randomly selected from the dataset. Hence, in the latter case, there is a session-wise overlapping between test and train sets unlike in the former partition.

3.4.1 Local Binary Patterns (LBP)

LBP is a commonly used image descriptor method used in computer vision. It transforms an image with more detail. Ojala et al [209] introduced this texture descriptor. Pixels in each 3X3 block in an image are compared with the central pixel. After thresholding, these pixel values are multiplied by powers of two and then added to get the value for the central pixel. Each pixel in the block has eight neighbourhood pixels. Hence, $2^8 = 256$ values are calculated based on the relative pixel value of the central pixel and the pixels nearby. Fig. 3.4 shows the different stages involved in LBP histogram extraction from an RGB image and how the extracted features differ for real and fake faces.

Include figure



(a) Real face (b) Gray-scale image (c) LBP (d) LBP histogram

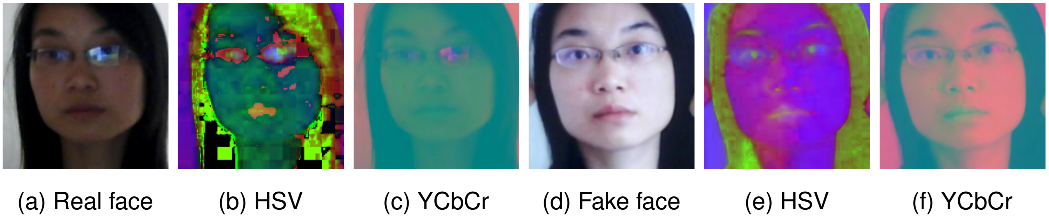
(e) Fake face (f) Gray-scale image (g) LBP (h) LBP histogram

Figure 3.4: LBP histogram for real and fake facial images

Support Vector Machine with linear kernel ('SVM-linear') and radial basis function kernel ('SVM-rbf') were used to classify client and imposter faces. Hyper-parameter tuning was done for each using Grid search. Predictions were made using the test set to determine model accuracy. Hyper-parameter tuning was carried out while testing with both partitions. In both test cases, SVM linear had the regularisation parameter, C=50 and gamma=.0001 as the best parameters. For SVM-rbf, C=50 and gamma=0.005 were the hyper-parameters.

3.4.2 Colour Texture Analysis

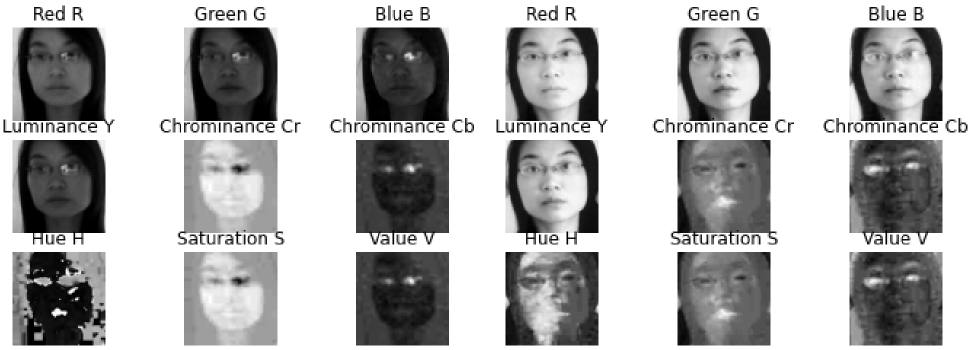
Spoofed images include local colour variations which are introduced during the recapturing process. In LBP extraction, the face images were converted to gray-scale images and ignored the colour texture variations associated with these images. To address this drawback, colour texture analysis was proposed [8] for face presentation attack detection. The real and fake facial images in three colour spaces, RGB, HSV and YCbCr are shown in Fig. 3.5.



(a) Real face (b) HSV (c) YCbCr (d) Fake face (e) HSV (f) YCbCr

Figure 3.5: Real and fake facial images in different colour spaces

Despite being the most commonly used colour space, luminance and chrominance information cannot be extracted properly from RGB. Hence, in order to utilise luminance and chrominance information to detect PAs, RGB image was converted to HSV and YCrCb colour spaces. After separating channel-wise components in each colour space, LBP histogram was calculated for each component. Once LBP histogram was calculated for all 6 channels (H,S, and V in HSV colour space and Y, Cr, Cb in YCrCb colour space), they were combined to form the resultant histogram. This histogram was then fed to a classifier to identify the deceived face image. Non-rotational invariant uniform LBP was used in the experiments. Images from each channel in the three colour spaces are shown in Fig. 3.6 for real and fake facial images.



(a) Real face

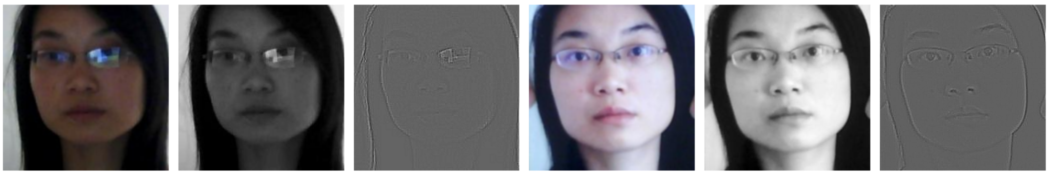
(b) Fake face

Figure 3.6: Channel-wise images from three colour spaces

3.4.3 Image distortion Analysis (IDA)

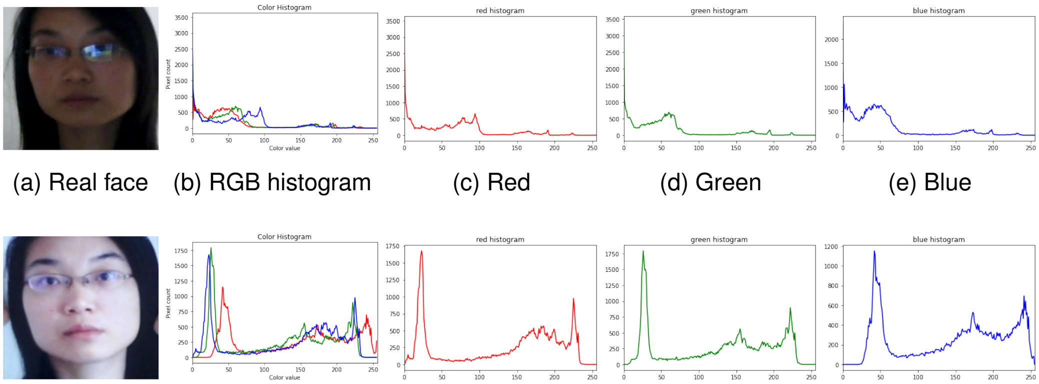
In this method, LBP was combined with image distortion analysis [197]. Three features were considered for analysing image distortion. They are colour moment, sharpness and colour diversity. Since fake face images have poor sharpness features, Laplacian filtering was applied to the image after converting it to gray-scale image. The Laplacian-filtered real image will have a high variance compared to the spoofed image. Hence, calculated variance from the face image would provide a cue to identify fake faces. Images after gray-scale and Laplacian convolution corresponding to real and fake facial images are illustrated in Fig. 3.8.

Similarly, real and fake images differ in colour distribution. This difference can be measured using colour moment. For these experiments, lower moments are calculated for each channel, after converting RGB image to HSV colour space. Since colour details loss happens in recapturing, colour diversity was also considered as a feature to detect PAs. The degree of colour diversity was calculated using choosing 10 pixel values which repeat the most in each channel of an RGB image. Thus 256 feature values from LBP, 1 from sharpness, 9 from the colour moment and 30 from colour diversity combined together to create a feature vector of 296 values as shown in Table. 3.2. However, Principal Component Analysis (PCA) was carried out to reduce the dimension of the feature vector to 150 and then the classifier was applied to this feature vector.



(a) Real face (b) Gray-scale (c) Laplacian (d) Fake face (e) Gray-scale (f) Laplacian

Figure 3.7: Real and fake images after Laplacian convolution



(a) Real face (b) RGB histogram (c) Red (d) Green (e) Blue

(f) Fake face (g) RGB histogram (h) Red (i) Green (j) Blue

Figure 3.8: Colour diversity in RGB histograms for real and fake facial images

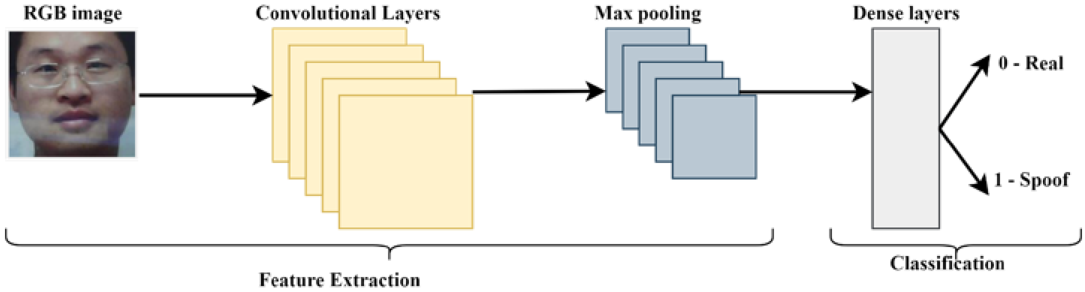
Face images and videos are displayed using varying printing materials and screens. Different spoofing mediums have different colour reproduction capabilities, and thus the colour distribution of the images displayed on them will also differ. As a result, these disparities in colour distribution can be used to detect PAS. Chromatic moments can therefore be used to analyze colour distributions. According to [210], three lower-order moments are calculated for each channel. To accomplish this, the RGB image is converted into HSV colourspace. A change in light does not affect the hue of an image. Furthermore, it is not affected by brightness, contrast, or white light reflection. Accordingly, chromatic moment features can be calculated by taking into account the average intensity, deviation, and skewness of each HSV channel.

Table 3.2: Dimensions of each feature vector in LBP and IDA

Extracted Feature from face images	Dimension of feature vector
LBP	256
Sharpness	1
Chromatic moment	9
Colour Diversity	30

3.4.4 CNN models

FPAD was carried out using both shallow and deep neural network models. The models were trained and tested for two different dataset partitions. The first partition had 3491 train images and 9123 test images, whereas, in the second partition, the entire dataset was divided into 80% train set and 10% each in validation and test set. Face images of size 64X64 from the NUAA dataset were resized to 150X150 pixel size. These two portions were utilised to evaluate five CNN models. The first model has three convolutional layers and one fully-connected layer. The second CNN model had four convolutional layers and three dense layers as in [197]. These two models were trained from scratch and results were reported.



(a)

Figure 3.9: CNN method for FPAD

In order to evaluate deeper model performance in FPAD VGG-16 [21] and ResNet-50 [20] were also considered. These pre-trained models were tested with both partitions. To achieve optimal performance, a suitable optimizer and learning rate were used while compiling the model. However, extensive hyperparameter tuning was not executed. Except for ResNet50, all the other CNN models used RMSprop optimizer with a learning rate=0.0001. ResNet50 used Adam optimizer with a learning rate=0.0001 for the 3491:9123 dataset partition, whereas for the other partition, ResNet50 used

Stochastic Gradient Descent (SGD), with a learning rate=0.001. Even though hyperparameter tuning was not carried out extensively, suitable optimisers were found out through experiments only. Models were compared based on their test accuracy.

3.5 Impact of customising dataset partitions in FPAD

The evaluation results of extracted features for two different dataset partitions, with SVM and RF are presented in Table.3.3 demonstrate the evaluation results of CNN models with both dataset partitions considered in the experiments. While comparing results of both manual feature extraction and CNN models, it is evident that when trained with 80% of dataset, both feature extraction and CNN showed the best results. LBP results are comparable with existing handcrafted feature methods for FPAD [200]. However, LBP performance is lower than all the CNN models considered in the experiments with this dataset partition. On the other hand, with 3491 training images, LBP showed almost equal or better results than CNN models.

Apart from LBP, all the other features showed reduced accuracy, when trained with the actual NUAA dataset partition. However, their performance had improved when trained with 80% data. These features showed best performance with RF classifier when trained with 80% data, but SVM-rbf gave best results with the actual partition. This implies that by properly managing training and testing images in a dataset, the performance of a model can be improved to achieve desirable generalisation.

Table 3.3: Performance of feature and CNN methods on different dataset partitions

FPAD Accuracy (%)		
Model	**80:20**	**3491:9123**
LBP+SVM-Linear	96.00	90.92
LBP+SVM-rbf	91.35	95.79
LBP+RF	98.45	69.30
CLBP+SVM-Linear	97.07	87.49
CLBP+SVM-rbf	98.18	87.94
CLBP+RF	99.18	68.98
LBP+IDA+SVM-Linear	97.83	63.41
LBP+IDA+SVM-rbf	97.97	75.52
LBP+IDA+RF	99.33	59.48
3 Conv+1 Dense	99.87	86.69
4 Conv+3 Dense	99.87	90.06
VGG-16	99.98	76.62
ResNet-50	99.64	80.03

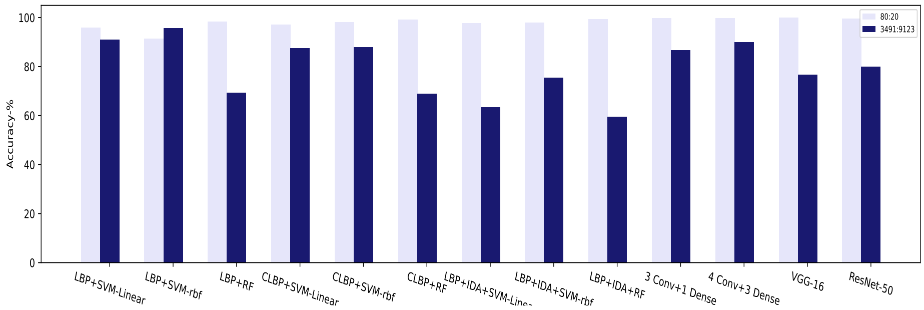


Figure 3.10: Performance comparison of models with two data partitions

The CNN models were trained with two different train-test ratios. It is evident from the Table. 3.3that with a training set having 3491 images, from the first two sessions had lesser accuracy than the training set having 80% dataset with random images from the dataset. The major reason for this result was in the dataset partition as in [10], the training set consisted of images from the first two recording sessions only. There are no images from the third session in the training set. On the other hand, the test set was formed using images captured in the third session. Hence there was no overlapping between the training and testing set. So, in the train and test ratio mentioned in [10] there would be less generalisable features common to both sets of images. Hence, it would reduce the accuracy of detecting a spoofed face.

In the second test scenario, the dataset is randomly divided in the ratio 80:10:10 as train, validation, and test sets. Since the whole dataset was considered for random selection, images from all three sessions would have been included in the training, validation, and test sets. Also, there were more training images in this dataset partition than in the earlier one. These two factors entirely altered the features, which were learned from training. The size and variance of the training set provided more generalisable features, which would facilitate better accuracy in identifying fake faces.

It can be seen in Table. 3.3 that VGG-16 and ResNet50 performance is lower compared to the other two shallow networks with 3491 images in the training set. The pre-trained network weights were derived from a model that was primarily used for image classification. Even though spoof detection is an image classification problem, some of the initial layer features in these pre-trained deeper networks may not be suitable for this specific task. This would result in some parameters which are not useful for the spoof detection problem. Hence, these deeper pre-trained models degrade in performance. However, with the 80:10:10 dataset partition, all the models detect spoofed faces almost perfectly, comparable to the existing state-of-the-art accuracy in intra-dataset evaluation [197, 199].

The performance is presented in Table. 3.3 was achieved in the intra-dataset evaluation of models with the NUAA dataset. However, in order to assess the generalisation capability of the model, cross-dataset performance should be evaluated. But existing models such as [199, 202] reported only intra-dataset performance unlike [197]. It would also be important to confirm the number of images used to train the model and their distribution, especially when the NUAA dataset was involved. Even though [10] provided a non-overlapping train and test sets for the NUAA dataset, recent research implemented models which did not follow that specific dataset protocol.

The NUAA dataset has features which facilitate training a generalised model. Even though images were captured in controlled settings, the session-wise capturing of images provided illumination variance similar to a real-life scenario. Hence, following the dataset partition in [10], aspects of the challenge of generalisation in FPAD can be addressed. However, in the existing literature, instead of developing the models to meet generalisation requirements with the NUAA dataset protocol, various train-test set ratios were used to increase model accuracy. The performance variations with two different partitions are presented in Table. 3.3 demonstrate this fact. Based on the exploratory experiments on the NUAA Imposter dataset and related research showed that the generalisation problem in FPAD was addressed by customising the dataset train-test ratio.

3.6 Conclusion

This chapter presents an experimental framework for detecting PAs that uses extracted features and deep learning. Based on the results of these experiments, CNNs trained on the original NUAA dataset partition achieve performance that is comparable to conventional feature extraction methods trained on the same partition. Using custom dataset partitions, even simple CNNs can achieve near-perfect accuracy, but this is likely not due to the CNN architecture, but rather to a more uniform distribution of data between training and test sets. This indicates that when proprietary dataset partitions are defined, data-hungry machine learning models can be supplied with sufficient samples. However, it also minimises challenges inherent within the NUAA dataset related to generalisation. In accordance with this analysis, a training dataset containing more attack variants will be compiled based on existing FAS datasets. Deep models will be trained using this aggregated dataset and the impact of data aggregation will be analyzed.

Chapter 4

Leveraging Data Aggregation for FPAD: An Empirical Approach

In this chapter, an experimental framework is presented to assess the impact of data aggregation on FPAD. As part of the experimental framework, it is examined how publicly available face anti-spoofing datasets may be aggregated to enhance the performance of cross-dataset evaluation. The generalisation ability of pre-trained deep models trained on the aggregated dataset was assessed using four popular and commonly used public datasets. The main findings of this chapter were published in the journal "Cognitive Computation" as "Deep transfer learning on the aggregated dataset for face presentation attack detection" [11].

4.1 Overview

Presentation Attacks (PA) are extremely diverse due to the wide variety of existing attacks and the unlimited potential for new ones to emerge in the future. Photo attacks have different variants, such as warped, printed, eye-cut and displayed photos [41]. Even within each variant of the attack, there will be differences based on domain-dependent features. These domain-dependent features include: capturing device, the material used for printing, illumination, resolution, display device, and the physical environment also causes variance. Video attacks have variants based on resolution and display devices. Manufacturers use distinct materials to make masks. Paper masks and wax masks are rigid masks, whereas silicon masks and rubber masks are non-rigid masks [7]. Flexible masks like silicon masks are much harder to detect because of their close similarity to human skin texture and appearance. Photo and video attacks are effortless to reproduce. The seamless access to personal images and videos through social media also helps in replicating 2D attacks with ease. So much diversity among PA techniques presents a significant challenge to PA detection methods. Successful Face Presentation Attack Detection (FPAD) needs to generalise across as many existing PA techniques as possible. In addition, these models should generalise across various domain-dependent features.

Although existing FPAD methods demonstrated impressive intra-dataset performances, they were not able to generalise against unseen attacks. These methods might have used available public FAS datasets for training [14]. These datasets have limited variance in terms of attacks and domain-dependent features. In contrast, attacks are more diversified in real-life scenarios. They differ in attack types and domain features. Hence, models trained on existing datasets may not generalise against such unseen attacks or even known attack formats in new physical environments. As a result, the reliability of the FR system deteriorates in practical applications. Moreover, emerging novel attacks have become a major threat to the generalisation capability of the existing FPAD methods. This has led to further investigation of the generalisation problem in FPAD [1]. One way to tackle the challenge of generalisation in FPAD is to produce a large and comprehensive dataset with many diverse attack variants, simply by aggregating existing datasets. Hence, the impact of data aggregation is investigated in this chapter, to address the generalisation of deep transfer learning models in the FPAD context.

4.2 FPAD using Aggregated Datasets

Presentation attack detection has attained significant improvement over the years, especially with CNN-based models. these models showed reduced generalisation capability against unseen attacks in real-life scenarios when compared with their benchmark statistics. The major cause for this deteriorated performance is the limited variance in training datasets. Hence, unseen attack detection across a wide range of attacks and across different datasets is still considered a challenging problem [1]. Public FAS datasets include only a few attack variants and domain-dependent features, such as illumination, settings, spoofing medium, and recording devices. Many of the existing FPAD models used one of these datasets for training. Hence, the models showed biasing towards the training dataset, exhibiting reduced generalisation against novel attacks.

There have been several studies exploring the concept of data aggregation to address the generalisation problem in FPAD. Costa et al. [14] proposed an aggregated dataset to provide more variance in terms of attack types, lighting, recording devices, and resolution. The authors combined ten public datasets to build the GRAD-GPAD (Generalisation Representation over Aggregated Datasets for Generalised Presentation Attack Detection). They also used a uniform protocol to evaluate the colour-based [208] and quality-based [91] models. This framework was further extended in [211] including demographic bias analysis and finer categorisation of PAs based on different factors such as resolution, spoofing medium and materials. This enhanced aggregated dataset mimicked more realistic scenarios. Thus the GRAD-GPAD and protocols facilitated the evaluation of the generalisation capability of the state-of-the-art methods. However, this grand dataset did not include multi-spectral datasets because of data incompatibility.

Saha et al. [84] also addressed domain generalisation using multiple datasets. The authors used four public datasets: Replay Attack [46], CASIA [41], OULU-NPU [165] and MSU-MFSD [43]. Three datasets were included in the training set, whereas the fourth was used for evaluation. The model learned the features from the three training datasets as single domain features. Thus, the model could use more domain dependent features, leading to better detection performance. Following the concept of dataset aggregation to improve domain generalisation, Nikisins et al. [91] combined three public datasets to illustrate the drawback of binary classification methods in detecting unseen PAs and evaluate their one-class classification model. The authors also established a specific evaluation protocol for the aggregated dataset, combining Replay Attack [46], Replay-Mobile [161] and MSU-MFSD [43]. The train, development, and testing sets were disjoint sets in terms of attacks. The aggregated dataset showed lower HTER (Half Total Error Rate) with Image Quality Measure (IMQ) methods when all the PA samples were part of the training set. However, binary classification exhibited poor performance on unseen attack detection. Authors of [193] used CASIA instead of Replay-Mobile to form an aggregated dataset. The authors of [212] and [213] used data aggregation in FPAD. Both of these works combined the real faces from the datasets, keeping the attack faces from each dataset with different domain features dispersed. They adapted this procedure to attain a generalised feature space.

Transfer learning utilises learned knowledge from one task for other similar ones. It assists in mitigating overfitting due to data limitations. Not only that, transfer learning saves computational resources as it avoids training deeper networks from randomised initial parameters. FPAD is a binary classification problem as it identifies if spoofing is present or not, and FPAD datasets are typically visible light spectrum, RGB images. Hence, a deep network that was trained for image classification with datasets like ImageNet [23] can be used to formulate a model to detect PAs. These pre-trained networks were used with fine-tuning either only top layers or a few convolutional layers with top layers.

In [75], Lucena et al. used transfer learning to address the FPAD problem. The authors fine-tuned a VGG-16 model that was pre-trained on ImageNet. Evaluation with a face spoof detection dataset demonstrated improved results over the existing the-state-of-the-art methods. Nagpal and Dubey [74] carried out extensive experiments using different pre-trained models to detect spoofed faces. The authors observed that transfer learning with deep models provided better results than using these networks with random weights or training from the beginning. Yu et al. [78] proposed a face anti-spoofing model using neural architecture search and transfer learning. In [214], the authors used transfer learning and Short Wave Infra-Red (SWIR) images for FPAD. A pre-trained face recognition network was used for transfer learning. Authors of [215] adopted a novel method to detect spoofed faces using extracted intrinsic image features and transfer learning. ResNet-50 [20] was used for implementing transfer learning which enhanced spoof detection using the extracted features from the datasets NUAA, CASIA and Replay Attack. Tu and Fang [216] utilised transfer learning using ResNet-50 and the Long Short-Term Memory (LSTM) to address FPAD. Compared to the state-of-the-art methods using feature extraction and shallow networks, these transfer learning-based methods exhibited better detection performance.

George et al. [38] used Light CNN, which is a pre-trained FR model and the concept of Domain Specific Unit (DSU) to address FPAD. This method utilized a multi-modal dataset with four modality. The low-level layers were re-trained using the new dataset and re-used the higher level weights. The extracted features from each modality data were concatenated together to form a final feature vector which was then passed to a fully connected layer of size 10 followed by sigmoid layer for classification. In this way a pre-trained FR model was fine tuned to adapt to the FPAD task using multi-modal data. Authors of [17] fine tuned the face recognition CNN model pre-trained on LWF [217] dataset, similar to aforementioned [38], to address domain adaptation of PAs in NIR. The initial two convolutional layers and first fully connected layers were made trainable in the fine-tuning. This facilitated the pre-trained model adaptation to the PAD task with NIR images. Even though the models were pre-trained on RGB data, the authors recorded a new NIR dataset with variance in illumination, environmental settings, subject pose, appearance and attack types. The model was able to detect photo and video attacks better than mask attacks.

The authors used the pre-trained FR model, Light CNN as a backbone/feature extractor to set up patch pooling concept to address FPAD, in [218]. Li et al. [219] proposed another dual mode method using NIR and RGB data to detect spoof. The authors used a light-weight network MobileNet-V3 as the backbone of the model. Each branch of the model was used to extract features from NIR and RGB data separately using this pre-trained model. The selected features were then passed to a softmax layer for classification.

Even though existing literature has explored the concept of combining multiple datasets for training the FPAD model, they rarely included the NUAA imposter dataset. Various handcrafted feature methods and deep learning methods were evaluated using NUAA. Either official or customised partitions were used to evaluate these methods [196]. Transfer learning and the concept of data aggregation were used to address the generalisation in FPAD. The experiments used a combined training set of official training partitions from NUAA, CASIA and Replay Attack. These three datasets have distinct 2D attack variants and domain-dependent features.

4.3 Deep Transfer Learning on the Aggregated Dataset

The experimental framework used transfer learning with binary classification to perform FPAD. Pre-trained deep networks, VGG-16 [21], ResNet-50 [20], Inception V3 [22] and DenseNet-121 [220] were used for transfer learning.

Three widely known, public datasets, NUAA [10], Replay Attack [46] and CASIA [41] were considered for these experiments and forming aggregated datasets. These three datasets and their combinations were used for training. All three datasets followed their official train/test split. Real face images from the three datasets combined to form a real face class in an aggregated dataset. Similarly, an attack class also was formed using attack images from these datasets. The combined train set provided different attack variants. For cross-dataset evaluation on this aggregated training set, SiW [102] test set was also used.

4.3.1 Datasets

The experiments used three FAS datasets. They were NUAA, CASIA-FASD and Replay Attack. In the existing literature, both traditional hand-crafted feature extraction methods and recent deep learning methods in FPAD have used these three datasets for evaluation. The different attack variants, test protocols and lighting conditions also assist in creating more variance within the aggregated dataset. CASIA and Replay Attack datasets, as distributed, consist of videos. Frames were extracted with a rate of 2 fps and face detection was carried out on these frames. NUAA was accessed as face-detected images, which are provided as part of the official dataset. These face-detected images were resized to 224 X 224 pixels. Official test/train partitions were used for each dataset. The experiments also used a SiW test set to perform the cross-dataset evaluation on a combined train set consisting of NUAA, Replay Attack and CASIA train sets. The facial images were extracted at a frame rate of 1 fps from each video to form this dataset. The SiW train set was unused. Table. 4.1 shows the number of train and test images in each dataset, which were used in the experiments.

Table 4.1: Number of train and test images in each dataset

Dataset	**Train**	**Test**
NUAA	3491	9123
Replay Attack	6950	7573
CASIA	5788	6469
SiW	40790	34779

The NUAA Imposter Database contains authentic images as well as photo attack and covers samples of 15 individual subjects. In contrast to the training set, the official test set is considerably larger. The training set contains 3491 images, while the test set contains 9123 images. These images were extracted from videos recorded at three different sessions under different lighting conditions. However, the already extracted images after face detection are available to the public. In NUAA's train partition, both classes have nearly the same number of images, whereas in the test set the attack images are much more numerous than the real face images. In terms of attack variants, it consists only of photo attacks. Despite these facts, NUAA remains popular among FPAD researchers [221, 222, 223].

CASIA-FASD has print attacks, warped photo attacks, cut photo attacks, and video attacks. 50 subjects were represented with fake and real faces. There are three real face videos and nine fake face videos for each subject. The train set features 20 people. There are genuine and fake videos of 30 individuals in the test set. The train and test sets are disjoint in terms of subjects. There are three real face videos and 9 attack videos corresponding to each subject. Thus, the train set has 60 real face and 180 attack videos in total. The test set includes 90 real face and 270 attack videos. Like NUAA, CASIA lacks ethnically diverse subjects. In addition, CASIA includes seven test cases, including three attack types, three image quality levels, and the entire dataset. In this experiment, the entire dataset is used with the given partition based on the dataset protocol. As this dataset has more attack variants, including video attacks, it is widely used for evaluating FPAD models [224, 225, 226].

Replay Attack was created by using 50 identities. There were respectively 15 subjects for training, 15 subjects for development, and 20 subjects for testing. While recording the Replay Attack dataset, printed images, mobile displays, and tablets were utilised. The three mediums were either fixed to a support or held by the operator during the recording process. Two types of recording environments were used to capture the videos, controlled and adverse. The controlled setting had a uniform background and illumination using incandescent lamps, whereas the adverse setting had a non-uniform background and day-light illumination. There are various PA types in this dataset. Hence it is popular among the FPAD researchers [227, 228]. Both train and development sets have 60 real face videos and 300 attack videos. The test set consists of 80 real face videos and 400 attack videos.

Spoof in the Wild (SiW) [102] dataset consists of 165 subjects from a more diversified ethnicity than the other datasets. There are 8 real face and 20 attack videos corresponding to each subject. Thus the dataset has 4,620 videos. The dataset was made using 6 spoofing mediums. Four different sessions were used varying factors such as poses, illuminations, expressions (PIE), and distance-to-camera. Videos were preprocessed by first using the frame rate to extract one image per second. Then the face area was extracted using the annotations provided. To increase diversity of facial images, the face area was cropped to accommodate some background information. This was achieved by multiplying each bounding box with a random scaling factor between 1.1 and 1.4. As with the other datasets, the images were resized to 224 X 224.

4.3.2 Aggregated Dataset

This experimental framework focuses on examining the impact of data aggregation on the generalisation of FPAD. To accomplish this task, an aggregate train set was constructed with NUAA, CASIA, and Replay Attack datasets' train partitions. NUAA consists of print attacks. Replay Attack has video attacks. CASIA includes both photo and video attacks. CASIA has warped, print, eye-cut photo attack variants. Thus the resulting aggregated train set has both video and photo attacks with variance in attack types and domain-dependent features.

Table 4.2: Number of real and fake images in three datasets and aggregated dataset.

Dataset	**Train**		**Test**	
	Real	Fake	Real	Fake
NUAA	1743	1748	3362	5761
Replay Attack	1689	5261	1928	5645
CASIA	527	1760	824	2471
Aggregated Dataset	3959	8769	6114	13877

The number of images in each class corresponding to three datasets and the combined dataset is shown in Table. 4.2. In Figure. 4.1, the distributions of real and fake classes in the individual and aggregated datasets are presented. NUAA has an almost equal number of real and fake class images in the train set. However, CASIA and Replay Attack have more fake face images than real ones in the train set. The aggregated train set includes 3959 real and 8769 fake face images.

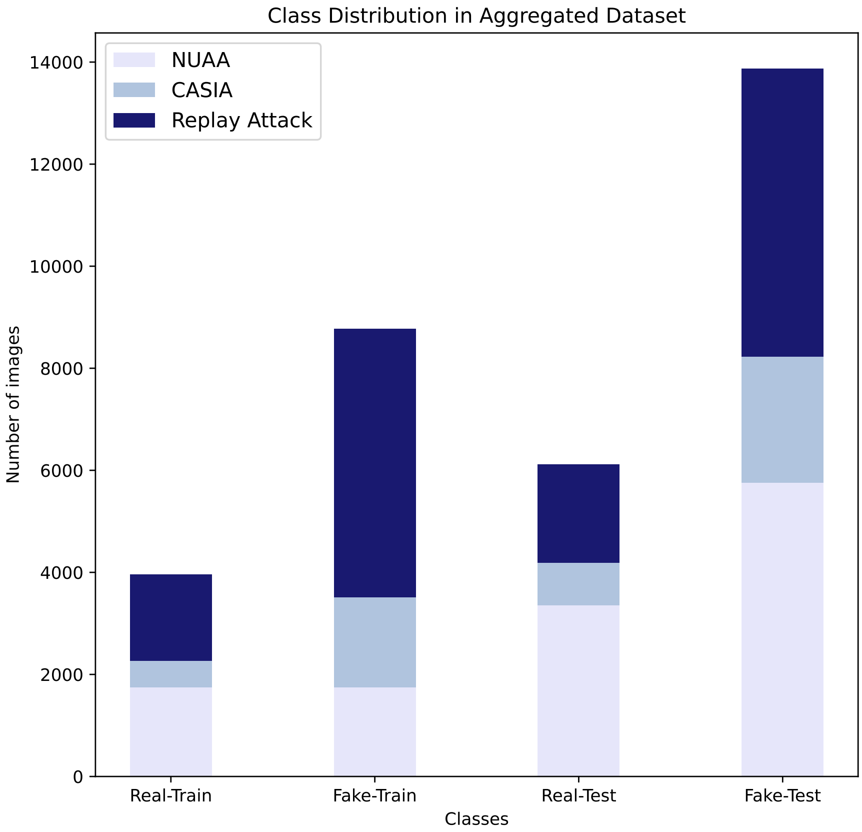


Figure 4.1: Class distribution in aggregated dataset

Given the unique characteristics of each dataset, such as lighting, the spoofing medium, the environment, and the recording device, the combination of these datasets produces greater variance in both real and fake classes. As a result, the model can learn a wider range of features to distinguish between real and attack classes. Furthermore, it avoids overfitting due to subtle biases within a single dataset. The individual training sets from each dataset were combined to form a training set, as indicated in Figure. 4.1. The aggregated test set was also constructed in this manner. Each dataset was divided following the official train/test protocol. Consequently, no mixing up of train and test set distributions occurred in the aggregated dataset. By keeping the distributions consistent with the official protocol, even in the aggregated dataset, we maintain the challenges of domain generalisation inherent to individual datasets [196].

4.3.3 CNN Models

An FPAD determines the authenticity of detected faces. In essence, it involves binary image classification. With deep CNN models, FPAD has also significantly improved, similar to any other computer vision task [75, 74, 221]. It must be noted, however, that deep neural networks require a substantial quantity of data to achieve desired performance. In order to solve such problems, transfer learning has become increasingly popular. By freezing some layers of the network and retraining others on a new dataset from the new domain, transfer learning re-purposes an already learned network to perform a similar task. In this manner, a task may be accomplished with less training data, less time, and higher accuracy. The majority of the FAS datasets are restricted in size. As a result, transfer learning was used to overcome this limitation.

The experimental framework utilises transfer learning to evaluate the aggregated dataset performance. For this purpose, we used pre-trained deep neural networks with architecture VGG-16 [75], ResNet-50 [216, 215], Inception V3 [74] and DenseNet-121 [108]. These models were popularly used and experimentally verified for FPAD in existing literature [221]. The networks used in the experiments were all pre-trained using ImageNet [23]. Pre-trained models were loaded without output layers, freezing the top layers. The models used "ImageNet" weights. The top layers were fine-tuned using FAS datasets to perform PA detection.

4.3.4 Experimental settings

The experiments included intra-dataset and cross-dataset evaluations using individual datasets and their different combinations. To carry out a cross-dataset evaluation on the aggregated train set, the SiW test set was used. Thus each model was evaluated using the dataset combinations as in Table. 4.3. Models were trained for 10 epochs with a batch size of 32. These parameters were the same for all classification models. A validation split of 20% of the train set was used while training the model. To compare the performance in binary classification ROC curve, accuracy, Half Total Error Rate (HTER), precision, recall, F1 score, False Positive (FP) and False Negative (FN) are used. The HTER is the average of the False Acceptance Rate (FAR) and False Rejection Rate (FRR).

As presented in Table. 4.2 and Figure. 4.1, two classes from three datasets were combined, and this aggregated train set was used to train the four models. For binary classification, there were real and fake classes irrespective of the actual dataset. The output layers in the base pre-trained model were replaced with one dense layer with a size of 1000 and a sigmoid layer. Binary cross-entropy was used as a loss function.

Adam optimizer [229] was used with all four models. The learning rate for VGG-16, ResNet-50 and DenseNet-121 was 10^{-5}. For Inception V3, the learning rate was 10^{-6}.

Table 4.3: Test set VS Train set combinations used in the evaluation

No.	Train Set	Test Set
1	NUAA	NUAA
2		Replay Attack
3		CASIA
4	Replay Attack	NUAA
5		Replay Attack
6		CASIA
7	CASIA	NUAA
8		Replay Attack
9		CASIA
10	NUAA+CASIA	Replay Attack
11	NUAA+Replay Attack	CASIA
12	CASIA+Replay Attack	NUAA
13	NUAA+CASIA+Replay Attack	Replay Attack
14	NUAA+CASIA+Replay Attack	CASIA
15	NUAA+CASIA+Replay Attack	NUAA
16	NUAA+CASIA+Replay Attack	SiW

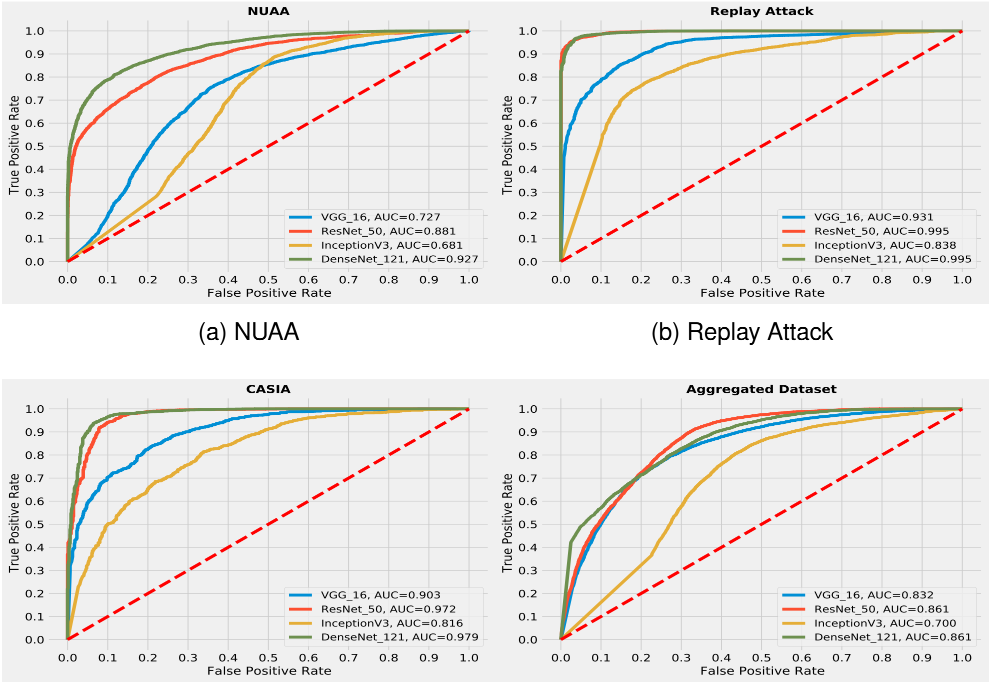
4.4 Impact of Data aggregation on FPAD

Extensive experiments and analyses were performed to investigate the influence of dataset aggregation in face presentation attack detection. Considered pre-trained models in the experiments were trained with three public FAS datasets and their combinations as in Table. 4.3. Both intra-dataset and cross-dataset evaluations were carried out to compare the model performance in FPAD. Intra-dataset evaluation results using individual and aggregated datasets are presented in Table. 4.4.

The Receiver Operating Characteristic curve (ROC) comparison of each model with all three datasets in the intra-dataset and cross-dataset evaluation scenarios is presented in Figure. 4.2 and Figure. 4.3. CASIA (93.35%) and Replay Attack (95.89%) showed the best intra-dataset performance with DenseNet-121 in intra-dataset evaluation. In contrast, NUAA had the highest performance (82.61%) with ResNet-50 in the intra-dataset evaluation.

Table 4.4: Comparison of intra-dataset and the aggregated dataset evaluations

Train	**Test**	**VGG-16**		**ResNet-50**		**Inception V3**		**DenseNet-121**	
		ACC(%)	HTER(%)	ACC(%)	HTER(%)	ACC(%)	HTER(%)	ACC(%)	HTER(%)
NUAA	NUAA	73.19	28.41	**82.61**	**19.08**	67.48	37.00	80.79	17.40
Replay Attack	Replay Attack	86.61	20.36	95.34	8.61	81.44	29.43	**95.89**	**6.76**
CASIA	CASIA	86.97	22.14	92.92	13.61	81.98	27.16	**93.35**	**12.85**
Aggregated Data	NUAA	**72.83**	**31.74**	71.18	38.72	67.97	40.40	64.61	42.71
Aggregated Data	Replay Attack	86.84	18.69	93.72	10.30	79.50	27.64	**97.32**	**4.37**
Aggregated Data	CASIA	86.45	20.16	92.02	11.66	81.69	28.83	**93.42**	**12.21**
Aggregated Data	Aggregated Data	79.981	25.770	83.177	26.574	74.584	34.100	81.447	25.528



(a) NUAA

(b) Replay Attack

(c) CASIA

(d) Aggregated dataset

Figure 4.2: ROC of Intra-dataset evaluation

The experimental framework also trained models with aggregated datasets. These models were then tested with individual test sets from CASIA, Replay Attack, and NUAA and with an aggregated test. In the aggregated dataset evaluations, NUAA datasets exhibited the lowest accuracy with more than 40% HTER (Table. 4.4. False Positive Rate (FPR) increased in the aggregated dataset evaluation on NUAA test sets. This increase in FPR caused lower accuracy and higher HTER for NUAA with all four model architectures. On the other hand, CASIA and Replay Attack showed a decrease in FPR in the aggregated dataset evaluation. ResNet-50 with Replay Attack was an exception, where FPR increased from 16.65% to 18.52%. As the FPR increased, it lowered the accuracy slightly. With DenseNet-121, both FPR and False

Negative Rate (FNR) decreased for Replay Attack in the aggregated dataset evaluation. These decreased FPR and FNR facilitated performance improvement in this specific evaluation. On the other hand, for CASIA, despite the decreased FPR, FNR doubled in the same evaluation scenario, resulting in only a slight improvement in accuracy (93.35% to 93.42%).

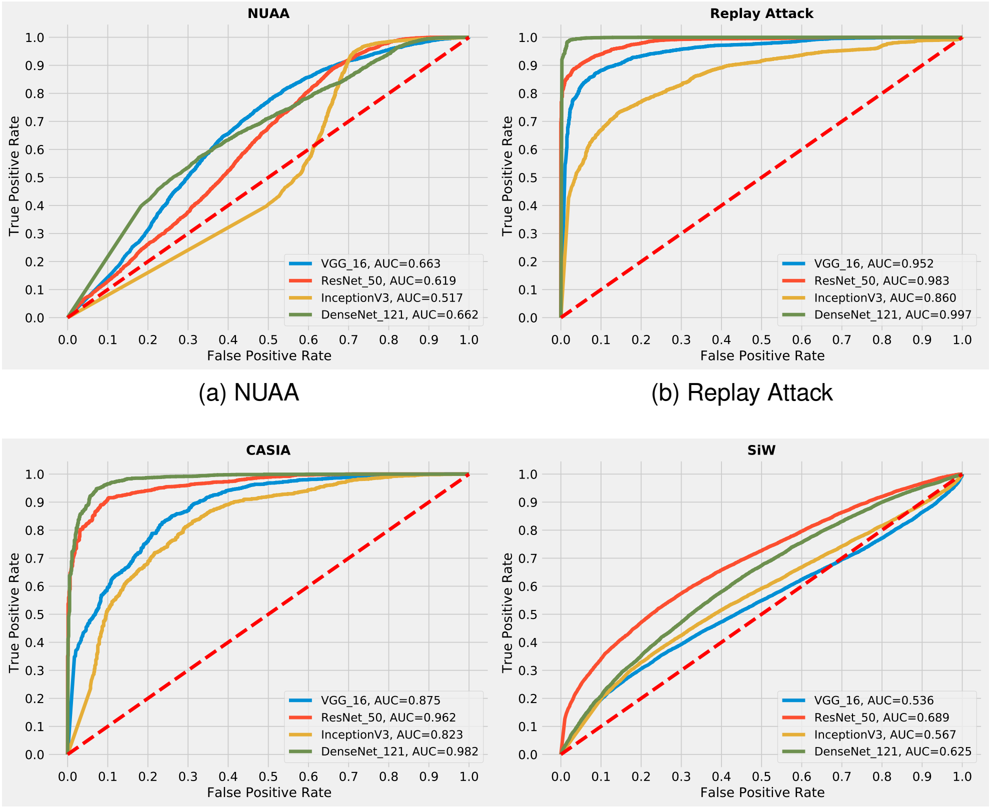
Cross-dataset evaluation results are shown in Table. 4.5. The evaluation was carried out using individual datasets and their combinations for training (Table. 4.3). To evaluate the performance of the aggregated dataset, the SiW test set was also used. The corresponding ROC is shown in Figure. 4.3 (d). It is evident from the plot that the cross-dataset performance of the aggregated dataset is significantly low compared to both intra-dataset performance and testing with other individual test sets. It would seem that SiW is different enough from NUAA, CASIA and Replay Attack that even an aggregate training set will not enhance generalisation to any great extent. The FPR in detection is more than 50% in most of the testing scenarios regardless of the datasets used, which caused higher HTER.

Table 4.5: Cross-dataset evaluation results

Train	**Test**	**VGG-16**		**ResNet-50**		**Inception V3**		**DenseNet-121**	
		ACC(%)	HTER(%)	ACC(%)	HTER(%)	ACC(%)	HTER(%)	ACC(%)	HTER(%)
NUAA	Replay Attack	49.95	50.46	56.68	49.81	32.13	45.54	54.38	41.10
	CASIA	53.78	41.45	54.48	44.35	54.78	33.43	71.08	30.12
CASIA	NUAA	68.39	34.07	59.05	52.94	69.30	37.53	59.37	50.38
	Replay Attack	69.51	48.66	67.42	52.14	50.44	47.47	73.17	50.32
Replay Attack	NUAA	57.71	52.24	59.26	53.08	63.38	49.68	61.93	50.96
	CASIA	30.05	49.03	58.03	57.67	74.48	50.14	68.47	53.18
NUAA+CASIA	Replay Attack	65.81	51.62	68.44	49.07	46.36	41.29	68.16	50.37
NUAA+Replay Attack	CASIA	66.04	40.64	46.56	51.57	60.36	43.82	68.13	45.23
CASIA+Replay Attack	NUAA	61.94	49.30	61.76	51.07	64.69	46.37	53.44	55.76
Aggregated Data	SiW	50.49	46.06	64.50	48.78	57.96	46.17	**62.87**	**38.79**

When trained with aggregated train set and tested with aggregated test set, the models had FPR more than 40%. This FPR value was much greater than the FPR value of testing scenarios with Replay Attack and CASIA test sets. Adding NUAA dataset while forming the aggregated dataset adversely effected the detection performance. This intra-dataset performance using aggregated datasets can be clearly demonstrated using the corresponding ROC, as in Figure. 4.2 (d). Unlike the intra-dataset evaluation on individual datasets, aggregated dataset performance diminished, even with DenseNet-121 (Table.4.4. For ResNet-50 and VGG-16, this aggregated data intra-dataset performance is near to NUAA intra-dataset performance. However, compared to the other two datasets, the overall intra-dataset performance using the aggregated dataset is low.

It is evident from Figure. 4.3 and Figure. 4.2 that, DenseNet-121 was the best model



(a) NUAA

(b) Replay Attack

(c) CASIA

(d) Aggregated dataset

Figure 4.3: ROC of models trained on aggregated train set against individual test sets

in both intra- and the aggregated dataset evaluations for CASIA and Replay Attack. However, it was ResNet-50 for NUAA rather than DenseNet-121. NUAA performed the best in the aggregated dataset evaluation with the VGG-16 model. It is evident from the plots that the performance on NUAA is not as good as the other two datasets in both evaluation scenarios. The cross-dataset performance evaluation using SiW dataset on the aggregated train set was even worse compared to testing the same model with other individual test sets as illustrated in Fig.4.4. All the models exhibited very low detection performance in this specific evaluation. This indicates that data aggregation alone does not help generalisation against various attacks.

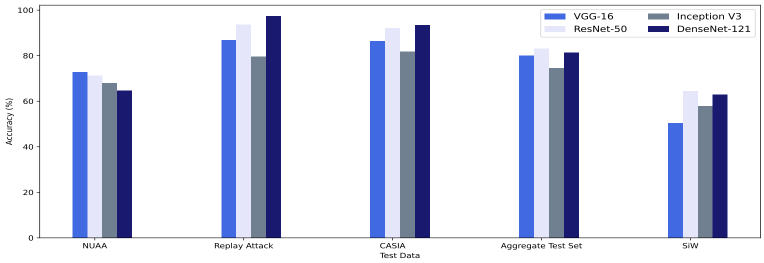


Figure 4.4: Graphical comparison of FPAD using data aggregation

4.4.1 Discussion

The aggregated dataset and cross-dataset evaluation results show that detection rates were reduced when tested with CASIA, Replay Attack, NUAA and SiW test sets when compared with models trained and tested on a single dataset. It clearly indicates that even though dataset attack variance and size improve with the aggregated dataset, it generally does not improve the detection performance on any component dataset. In fact, combining these datasets led to an increased FPR. Training with these combined datasets restricts the models from identifying real faces correctly. FPAD relies more on spoofing patterns and image quality features. As NUAA was recorded in 2010 [10], using a webcam, the image quality is lower compared to other datasets in the experiments. Similarly, CASIA also has images of three different qualities, including lower-quality images. This quality variation in images influences the high-frequency feature extraction while training the model. Regarding transfer learning, the deep networks used in the experiment were pre-trained for image classification tasks. They extract deep, global features. However, FPAD may require shallow, local features to detect spoofing. These pre-trained image classification models might have failed to learn spoof-specific features to achieve better detection performance, instead relying on some dataset-specific features.

In all the evaluation scenarios, false positives were more significant than false negatives. This shows that even though attacks were detected, the models failed in identifying the genuine images, particularly those in NUAA. This influences the overall performance of these models. CASIA and Replay Attack facial images were extracted from raw videos using the same pre-processing methods. NUAA is available to the public as pre-processed face-detected images. These images were resized for experiments. This disparity between the NUAA dataset and the other two dataset images may have influenced the classification performance.

The classification was carried out using four deep networks with different architectures. However, except for DenseNet-121, the other three models exhibited the same performance trend in the aggregated dataset evaluation scenario: the models trained and tested on the same datasets performed better than models trained on an aggregate dataset. Even DenseNet-121 followed the same trend with the NUAA dataset. This implies that combining source domains solely cannot improve the detection performance. The cross-dataset evaluation results are presented in Table. 4.5 support this.

With the combination of more datasets, handcrafted features were evaluated for generalisation capabilities [14] within the context of FPAD. Based on the results of this research, it was found that state-of-the-art methods with impressive intra-dataset performance are less generalisable in cross-dataset evaluation when used with a combination of heterogeneous sources. A variety of factors influence their performance, including their capture devices, display conditions, and image quality. In contrast to this evaluation, the experiments used binary classification using four pre-trained deep neural networks to detect PAs. It was evident from the analysis of experimental results that even deep learning frameworks were not capable of generalising to different distributions.

4.5 Conclusion

Data aggregation was used in this chapter to detect PAs using deep transfer learning models. When the models were trained with the aggregated training set and tested with test partitions from individual datasets, detection performance was lower than intra-dataset evaluation. In view of this, it can be concluded that combining multiple source domains alone is not sufficient to guarantee domain generalization against unseen attacks. To generalise FPAD, a method must be crafted in such a way that generalisable features can be extracted. Hybrid fusion methods have been used recently to extract and classify such features in the FPAD context. In the following chapter, a hybrid approach will be presented where deep features and colour texture will be combined to enrich the feature space.

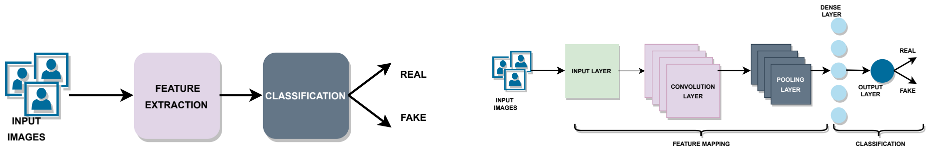
Chapter 5

Enhancing FPAD: Fusion of Color Texture and Deep Features

Based on the findings in the preceding chapter, it becomes apparent that achieving generalisation in FPAD necessitates specialised techniques that extract features capable of generalisation. Therefore, this chapter introduces an approach that combines deep and colour texture features. Thorough experiments unmistakably demonstrate the advantages of expanding the feature space to enhance detection rates. Additionally, this chapter includes a comparative analysis of the computational speed between the baseline method and the fusion method. This research paves the way for future investigations into enhancing Face Presentation Attack Detection (FPAD) by exploring novel features and fusion strategies. The main findings of this chapter were published in the journal "Sensors" as "Fusion Methods for Face Presentation Attack Detection" [12].

5.1 Overview

Traditional FPAD methods utilized manually extracted features such as texture, image quality, and motion, combined with standard classifiers, such as SVM and Random Forest, to determine whether the detected facial image is real or not [10]. Convolutional neural networks (CNN) took the place of these classical feature engineering models. Hand-crafted feature-based FPAD methods are shown in Figure 5.1 a. Figure 5.1 b shows deep learning-based FPAD methods. CNN-based FPAD models benefited from their exceptional inherent feature-extraction capability to some extent. Yet these deep learning-based models failed to reach adequate generalisation against emerging, unseen attacks [230].



(**a**) FPAD using handcrafted features (**b**) FPAD using deep learning

Figure 5.1: Face presentation attack detection (FPAD) methods

There are multiple reasons for the low generalisation capability of FPAD models. The majority of FPAD models were either designed for the detection of specific types of attacks or were trained by using the existing face anti-spoofing (FAS) datasets. However, these FAS datasets have limited variance in size, attack types, and subjects. Moreover, datasets were recorded in a controlled environment that lacked sufficient variation in illumination, recording devices, settings, and the environment [14]. As a result, even if these models detect some specific attack types, they are not reliable in detecting unseen attacks in real-life scenarios. This necessitates the development of more generalised FPAD models to detect PAs [1].

Some recent efforts to improve FPAD have leveraged features from models pre-trained on large datasets designed for object recognition [75, 74]. These datasets have high variance across multiple factors. This led to the models performing well in object detection, recognition, and captioning tasks that incorporated deep features from the images. The spoof detection problem does not have large, labelled datasets, unlike these computer vision tasks. Detecting presentation attacks involves detecting spoof-specific features, such as specular reflection, deformations, glare, spoof patterns, and Moire effects [226]. These features are not always present in high quantities in the common datasets designed for image-classification tasks. Hence, relying on deep models, which were pre-trained on image classification datasets, when the data does not exhibit the necessary features, may not be optimal for improving FPAD performance. Meanwhile, traditional feature extraction methods make use of shallow features. In this approach, the challenge is to select a suitable descriptor that is invariant to factors such as illumination, light, skin type, recording device, and environment. These descriptors should also effectively represent the spoof-specific patterns [230].

PAs, especially 2D attacks, are either printed on different materials or displayed on digital devices [5]. Mask attacks also can be created by printing the genuine face on suitable materials [7]. Such recapturing processes introduce distortions in PAs. The distortions are the cues to distinguish between real and fake faces [43, 231]. Texture methods (LBP, HOG, and DOG) were used to extract these cues for PA detection. A good number of texture-based methods used grayscale images, discarding colour feature-related cues. However, colour distortion cues provide significant information for identifying PAs [208, 8]. Hence, colour texture analysis was considered in this work to combine with deep features to perform PA detection.

FPAD performance has been significantly improved by hybrid methods in recent years. Section. 2.7 provides an overview of current hybrid methods. Thus, considering the advantages of hybrid methods, a hybrid fusion method has been proposed for addressing FPAD. This fusion method takes advantage of both handcrafted and deep features. Colour texture features, which provide spoof-specific cues, were combined with deep features extracted by using pre-trained image classification models. As a result, both the local features and the deep global features were used together as the input to the classifier to determine the authenticity of the facial images. Even though deep feature extraction through transfer learning, colour texture analysis, and their fusion for FPAD are already established methods in the related literature, those methods used traditional machine learning classifiers such as SVM. The fusion method presented in this chapter takes advantage of neural network-based classifiers. Instead of using any of the commonly used pre-trained models, this study compares the fusion method by using three commonly used pre-trained models and a custom CNN model trained from random initialisation.

5.2 Hybrid Fusion Method for FPAD

An experimental framework is used to detect PAs by fusing deep and hand-crafted features. For the evaluation, three publicly available datasets, CASIA [46], Replay Attack [41] and SiW [102] were used. For this fusion method, texture was extracted from the images by using colour texture analysis (CLBP) [8]. By using pre-trained deep learning models, VGG-16 [21], ResNet-50 [20], and Inception V3 [22], deep features were extracted. These high-level features from deep models and low-level features from colour texture analysis were then concatenated and passed to the classifier. The classifier consisted of a dense layer with 512 units and a sigmoid layer.

Additionally, a custom CNN model was trained only on each dataset individually in order to compare with the pre-trained networks. The resultant features were combined with the colour texture features and passed to a classifier as before. The fusion method used deep features from pre-trained and custom CNN models in different evaluation scenarios to compare the impact of fine-tuning and fully training models on FAS datasets. The experiments also consisted of baseline methods. The pre-trained and custom CNN models were trained for binary classification. As in the fusion methods, the baseline classifier also consisted of a dense layer with 512 units size and a sigmoid layer. All the evaluation scenarios used binary cross entropy as the loss function and Adam as the optimiser.

5.2.1 Pre-Trained Models

FPAD is conventionally treated as a binary image classification problem. Hence, FPAD also takes advantage of transfer learning to address dataset limitations. VGG-16 [75] and ResNet-50 [215] have been used previously to address FPAD by using transfer learning. Lucena et al. [75] fine-tuned the VGG-16 model by changing the top layers for detecting PAs by using binary classification. Nagpal and Dubey [74] used Inception-V3 and ResNet-50 models for the PA detection. According to the authors, transfer learning with these pre-trained models facilitated better detection performance than training from a random weight initialisation. The authors of [215] also used ResNet-50 for the FPAD task.

Pre-trained models, VGG-16, ResNet-50, and Inception V3 were used for binary classification as well as feature extraction in the presented experimental framework in this work. These deep network models were pre-trained for image classification [23]. The features were extracted by removing the output layer from the models. VGG-16 feature vector size was 4096. Feature vectors of size 2098 were extracted from ResNet-50 and Inception V3. For binary classification, the top layers were replaced with a fully connected layer of size 512 and a sigmoid layer in these pre-trained models. Thus, transfer learning was applied in the baseline methods.

5.2.2 Convolutional Neural Network (CNN) Model

Evaluation was also performed by using a custom CNN model. The model has five convolution layers, each followed by a max-pooling layer. The classification block in this model is formed by using a dense layer of size 512 and a sigmoid layer (Figure 5.2). From block 1 to block 5, the number of filters varied from 32 to 512. A kernel size of 3 $\times$ 3 was used in each convolutional layer. The models were trained by using a corresponding FAS dataset used in the experiments. The weights from these models were also used for feature extraction in the fusion method.

Similar to pre-trained models, the custom CNN model was also used to extract deep features by removing the output layer. This provided a feature vector of size 512. Compared to the deep models, VGG-16, ResNet-50, and Inception-V3, this CNN architecture was shallow, with only 8 layers. The CNN model used for feature extraction was shallower compared to the pre-trained models.

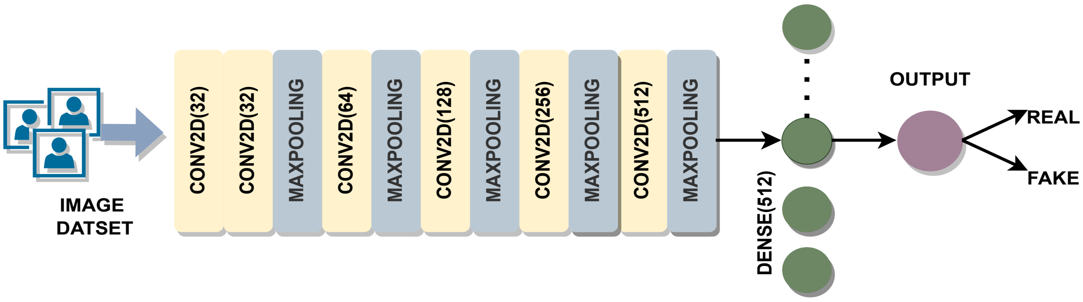


Figure 5.2: Custom CNN Model

5.2.3 Colour Texture Analysis

Presentation attacks include photos printed on different mediums, video or photo displayed on digital devices, and masks. The spoofing medium varies in resolution and display quality. Grayscale image-based texture analysis facilitates identification of high-quality PA. The grayscale-based methods (e.g., LBP) cannot provide sufficient difference in textural cues when the quality of the PAs diminishes [208]. A PA image or video passes through at least two cameras and a printing or display medium. Hence, many PAs have a recapturing effect. Compared to an authentic capture, the colour reproduction of these spoofing mediums would be limited. Hence, PA will have the colour features corresponding to the printing or displaying medium gamut. Moreover, the recapturing camera and the entire recapturing process to perform PA can cause colour disparities and imperfections.

Human eyes are more sensitive to luminance than chrominance. Hence, the colour reproduction mapping in printing or display process preserve luminance variation in the source image rather than chrominance. Thus, the PAs may contain chrominance variations which are largely invisible to human vision. These chrominance variation cues can be utilised to distinguish between real and fake facial images. The majority of the available FAS datasets provide RGB images or videos. On the other hand, RGB colour space has high correlation between the colour components. The RGB colour space does not adequately separate luminance and chrominance information. Recapturing introduces chrominance variation in PAs, while sustaining luminance variation. It is unlikely that RGB colour space would be able to determine spoof-specific chrominance cues. Thus, alternative colour spaces should be used to extract such discriminatory cues [8].

By analyzing the chroma channel colour texture, the local colour disparities discussed above can be identified. Luminance and chrominance are represented in YCbCr colour space. The chrominance component of YCbCr reveals disparities which are presented in PAs. HSV colour space represents hue, saturation, and brightness. HSV colour space contains a chrominance component, which is complementary to that in YCbCr colour space. Both of these color spaces provide chrominance components that can be used to identify PA. Hence, colour texture analysis was used to extract the hand-crafted feature in this proposed fusion method. Although deep networks provide global features, extracted features include the local (chroma) cues. Colour texture analysis [8] addresses these variations, extracting LBP from individual channels from the images. Hence, RGB images were converted to HSV and YCbCr colour spaces to extract related features to identify PAs. Figure 6.4 presents the process of colour texture analysis. To extract the colour texture features, the channel-wise components were separated after conversion to each colour space. An LBP histogram for each channel was calculated. The histograms from these 6 channels (H, S, and V in HSV colour space and Y,Cb,Cr in YCrCb colour space) were then combined to form the final feature vector of size 354.

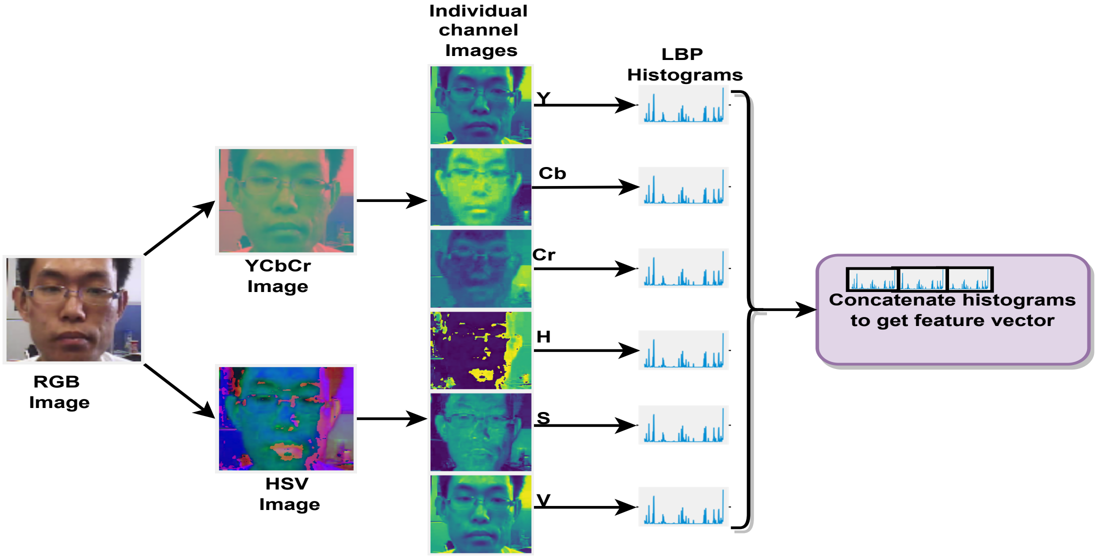

Individual channel Images
LBP Histograms
Y
Cb
Cr
H
S
V
YCbCr Image
RGB Image
HSV Image
Concatenate histograms to get feature vector

Figure 5.3: Colour texture analysis

5.2.4 Fusion Method

The experimental framework extracted and combined high-level deep features and lo- level local color texture features. Deep features were extracted by using transfer learning and custom CNN models. VGG-16, ResNet-50, and Inception V3 were used to get features' vectors of size 4096, 2048, and 2048 respectively. As mentioned in Section 5.2.1, these models were pre-trained on ImageNet. On the other hand, custom CNNs were trained on three FAS datasets and used to extract feature vectors of size 512. Because the RGB image was converted into HSV and YCbCr colour spaces, there were six channels in total. As the texture feature extraction using LBP was carried out on each channel, these channels provide a feature vector of length 59. Thus feature vectors from these six channels forms a low-level feature vector of size 354.

Concatenation is an effective way to combine different features for use in machine learning. Extracted CLBP features were concatenated with features either from pre-trained models or a custom CNN model. Thus, by concatenating sets of deep and handcrafted features, a final feature vector was created. Let F^{Deep} be the deep feature vector with size m and F^{CLBP} be the colour texture feature vector with size n. Then the final feature vector F^{Fusion} can be represented [16] as

$$F^{Fusion} = F^{Deep} \cup F^{CLBP}.$$

Thus F^{Fusion} will have the size $(m + n)$. Here F^{CLBP} had size 354. m, the size of F^{Deep} feature vector, held different values according to the pre-trained or custom CNN models, which was used for feature extraction. Hence the size of F^{Fusion}, m was determined based on the deep model used for feature extraction.

Figure 5.4 illustrates the structure of the proposed framework. It consisted of a deep feature extraction module and a hand-crafted feature extraction module. The resultant feature vectors from these modules were the input of the fusion module. Fusion simply involved concatenation. This combined feature vector was then passed to the classifier. The classifier consisted of a dense layer of size 512 followed by a sigmoid. However, the input size of the classifier was different according of the different feature vector size.

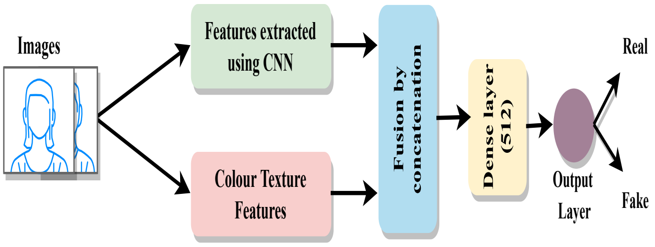


Figure 5.4: Fusion Method

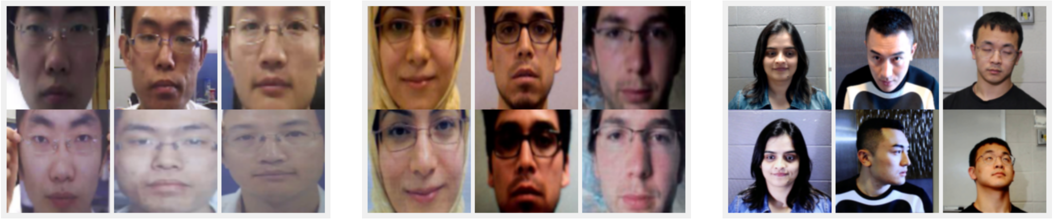
The final fused-feature classifier present in all models consisted of a dense layer of size 512 followed by a sigmoid.

5.3 Experiment

This experimental framework evaluated baseline models and proposed a fusion method using three FAS datasets, namely CASIA, Replay Attack, and SiW. The pre-trained models were fine-tuned to carry out PA detection. The customised model was trained by using the aforementioned three FAS datasets. The datasets and experimental settings including hyper-parameter tuning are explained below.

5.3.1 Datasets

By using three widely used public FAS datasets, CASIA, Replay Attack, and SiW, the proposed fusion method was evaluated. These FAS datasets mainly include print and replay attack PA types. The datasets differ from each other in terms of size, gender ratio, ethnicity, recording devices, spoofing medium, illumination, and settings. Figure 5.5 shows both real- and fake-face samples from three datasets. The top row in Figure 5.5a–c has genuine facial images. The lower row shows the corresponding fake facial images.



(**a**) CASIA (**b**) Replay Attack (**c**) SiW

Figure 5.5: Real and fake samples from the datasets

CASIA has fake- and real-face videos from 50 subjects. The dataset contains photo attack variants including print, warped photo, and cut photo. Video attacks are also part of this public dataset. Corresponding to each subject, there are 3 real-face and 9 fake-face videos. Thus a total of 600 videos are included in the CASIA dataset. The training set is made of videos of 20 subjects. The remaining videos from 30 subjects are included in the test set. The train and test sets are disjointed in terms of subjects. CASIA has videos of three qualities; low, normal, and high. This dataset was recorded in natural scenes. This process did not use any kind of artificial unification while recording the dataset. Because cut photo attacks were included in CASIA, eye blinking was also incorporated in the videos. Print attacks with better quality were reproduced by using copper papers. However, CASIA includes only subjects from a single ethnicity, i.e Asian. Even though 10% of the subjects were females, the training partition is devoid of female subjects. Hence there is no gender variance in the training partition.

In the Replay Attack dataset, videos from 50 subjects are distributed among training, development, and test sets. The training and development sets have 15 subjects each, whereas the test set has 20 subjects. Fake faces were displayed as printed photos, on mobile and tablet displays. During the recording process, these three mediums were either fixed or held by the operator. Both controlled and uncontrolled settings were used. Uniform background with illumination using incandescent lamps was used in controlled settings. An uncontrolled setting made use of non-uniform background and natural light for illumination. The Replay Attack dataset also had a female-to-male subject ratio 1:9. However, unlike CASIA, this dataset has both gender in all the data partitions. This dataset also has variance in terms of ethnicity.

SiW is the third dataset used for the evaluation of fusion method. Compared to the other two public datasets used for experiments, the SiW dataset includes variance in terms of ethnicity, poses, illuminations, expressions, and distance-to-camera. A total of 4620 videos from 165 subjects include 8 real face and 20 attack videos corresponding to each subject. A total of 27% of the subjects in SiW datasets are females. It has subjects belonging to different ethnicity: 35% Asian, 35% Caucasian, 7% African American, and 23% Indian people are included in the dataset, giving much more variance in ethnicity. Among the subjects, 44% have glasses and 20% have beards. Unlike the CASIA and Replay Attack datasets, which have only a frontal pose range, SiW has pose ranges of $[-90, 90]$. Moreover, SiW used artificial illumination. Table 6.1 shows a comparison of the three datasets in different aspects.

Table 5.1: Comparison of FAS datasets used in the evaluation of the fusion method.

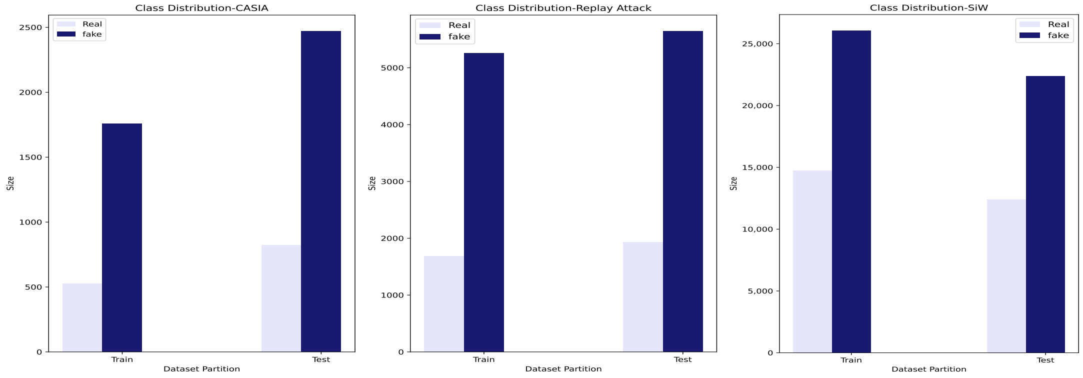
Dataset	subjects	Live videos	Attack videos	Attack types	Display devices
CASIA	50	150	450	2 Print, Replay	iPad
Replay Attack	50	200	1000	Print, 2 Replay	iPhone 3GS, iPad
SiW	165	1320	3300	2 print, 4 Replay	iPad Pro, iPhone 7, Galaxy S8, Asus MB168B

5.3.2 Experimental Settings

Three FAS datasets, CASIA, Replay Attack, and SiW were used to evaluate the fusion method. These datasets are available in video format. Frames from CASIA and Replay Attack were extracted at a rate of 2 fps and faces from these frames were detected. For the SiW dataset, frames were extracted at 1 fps and, by using given annotations, face detection was carried out. In addition, some random scaling of the bounding box for SiW was performed in order to provide some background information and improve facial image diversity. Facial images from three datasets were resized to 224 × 224 pixels. Rather than using customised data partitions to address generalisation [196], the official train-test split was maintained. Table 5.2 shows the number of training and test images in each dataset. Figure 5.6 shows the data distribution corresponding to each dataset.

Table 5.2: Sample size of each dataset in train and test partitions

Dataset	CASIA		Replay Attack		SiW	
Class	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**
Real	527	824	1689	1928	14733	12390
Fake	1760	2471	5261	5645	26057	22389
Total	2287	4118	6950	7573	40790	34779



(a) CASIA (b) Replay Attack (c) SiW

Figure 5.6: class distribution of the datasets after image extraction

Intra-dataset evaluation was conducted for the baseline and fusion methods. Images of 224×224 size were used in the evaluation. In the baseline method, the training used 10 epochs and 32 batch size with CASIA and Replay Attack. Training with SiW was carried out with a batch size of 512. The Adam optimiser [229] with different learning rates were used in the experiments. VGG-16 as well as ResNet-50 used a learning rate 10^{-5} whereas Inception V3 used 10^{-6}. A customised CNN model used different learning rates while training with different datasets. Learning rates of 10^{-6}, 10^{-5}, and 10^{-4} were used while training with CASIA, Replay Attack, and SiW respectively. The Python programming language was used for implementing experiments by using Keras with TensorFlow on the backend. Experiments were conducted on NVIDIA DGX-1 machine, using a single GPU system. The results are reported in terms of accuracy, half total error rate (*HTER*), precision, recall, F1 score, false positive rate (*FPR*) and false negative rate (*FNR*). *HTER* is the average of *FPR* and *FNR* (Section. 2.10).

5.4 The Impact of Hybrid Feature Fusion on FPAD: An Analysis

Fusion and corresponding baseline methods were evaluated by using CASIA, Replay Attack, and SiW datasets. For deep feature extraction, pre-trained models, VGG-16, ResNet-50, and Inception V3 were used. Custom CNN models trained on the three FAS datasets were also used for deep feature extraction. The fusion method combined deep features from each model with CLBP features and then classified by using a neural network-based classifier. This classifier had an input layer, one dense layer of size 512 units, and a sigmoid layer. Evaluations were carried out to compare PA detection accuracy, HTER, and computational speed for both baseline and fusion methods.

PA detection performance of baseline models are presented in Table 5.3. Binary classification using pre-trained and custom CNN models were considered as the baseline methods. Table 5.4 represents the results of fusion methods. A graphic representation of accuracy comparison is also presented in Figure 5.7. Figure 5.8 shows the ROC curve corresponding to baseline and fusion models.

Transfer learning using ResNet-50 had the highest accuracy among the baseline models for CASIA (93.36%), Replay Attack (95.57%), and SiW (98.78%). A custom CNN model performed better than VGG-16 and Inception V3 with Replay Attack, and SiW. However, with CASIA, all three pre-trained models had better detection rate than the custom CNN model in baseline evaluation (Table 5.3). However, ResNet-50 (98.78%) and custom CNN (98.16%) models exhibited very close detection accuracy in baseline evaluation with SiW.

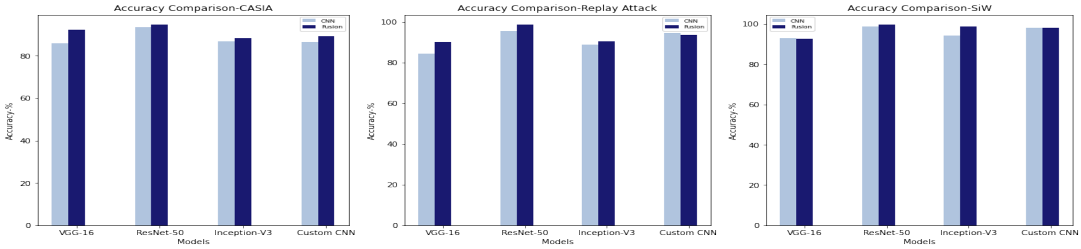
Table 5.3: FPAD results using deep CNN models

Dataset	Model	ACC (%)	HTER(%)	Precision	Recall	F1score	FNR(%)	FPR (%)
CASIA	VGG-16	85.85	24.01	0.87	0.96	0.91	4.29	43.69
	ResNet-50	**93.36**	**12.60**	0.92	0.99	0.96	0.69	24.51
	Inception V3	86.74	24.39	0.86	0.98	0.92	2.10	46.60
	Custom CNN	86.42	14.84	0.94	0.88	0.90	12.34	17.35
Replay Attack	VGG-16	84.25	23.05	0.88	0.92	0.90	8.18	37.91
	ResNet-50	**95.57**	**8.07**	0.95	0.99	0.97	0.66	15.51
	Inception V3	88.78	19.94	0.88	0.98	0.93	2.18	37.71
	Custom CNN	94.39	6.75	0.97	0.96	0.96	4.43	9.23
SiW	VGG-16	93.02	7.92	0.94	0.95	0.95	4.33	11.04
	ResNet-50	**98.78**	**1.57**	0.98	1	0.99	0.38	2.75
	Inception V3	94.35	6.53	0.95	0.97	0.96	3.48	9.57
	Custom CNN	98.16	2.24	0.98	0.99	0.99	0.85	3.63

From the fusion method results in Table 5.4, it is evident that the detection improved for almost all the datasets, regardless of the models used for deep feature extraction. Thus fusing colour texture features with deep features improved PA detection in intra-dataset evaluation. Among the models used, the combination of colour LBP (CLBP) with ResNet-50 features showed the best performance (Table 5.4).

Table 5.4: Fusion methods results.

Dataset	Model	ACC (%)	HTER(%)	Precision	Recall	F1score	FNR(%)	FPR(%)
CASIA	VGG-16+CLBP	92.33	13.26	0.92	0.98	0.97	2.06	24.39
	ResNet-50+CLBP	**94.65**	**8.68**	0.95	0.98	0.96	1.74	15.29
	Inception V3+CLBP	88.31	18.54	0.90	0.95	0.92	4.82	32.28
	Custom CNN+CLBP	89.34	15.47	0.92	0.94	0.93	5.87	25.12
Replay Attack	VGG-16 + CLBP	90.18	15.14	0.92	0.96	0.94	4.30	25.99
	ResNet-50 + CLBP	**98.56**	**2.64**	0.98	1.00	0.99	0.19	5.08
	Inception V3 + CLBP	90.38	16.19	0.91	0.97	0.94	2.82	29.56
	Custom CNN + CLBP	93.64	8.51	0.96	0.96	0.96	4.13	12.91
SiW	VGG-16 + CLBP	92.65	8.47	0.93	0.95	0.94	4.59	12.34
	ResNet-50 + CLBP	**99.60**	**0.48**	1.00	1.00	1.00	0.20	0.76
	Inception V3 + CLBP	98.61	1.57	0.99	0.99	0.99	0.95	2.20
	Custom CNN + CLBP	97.96	2.42	0.98	0.99	0.98	1.13	3.70



(a) CASIA

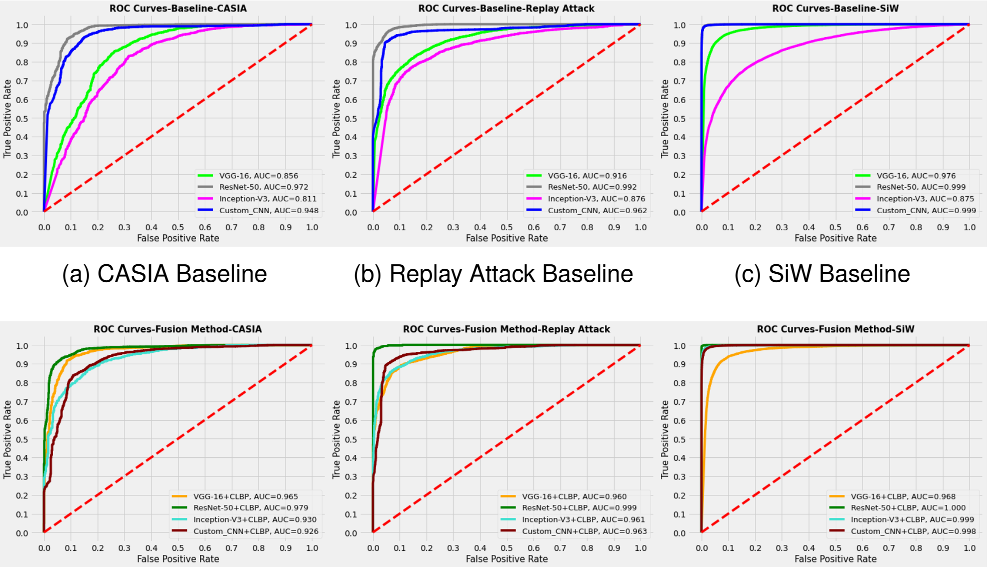
(b) Replay Attack

(c) SiW

Figure 5.7: Accuracy comparison for CASIA, Replay Attack, and SiW

A graphic comparison of detection performance of pre-trained and custom models is shown in Figure 5.7. It shows that fusing CBLP features with CNN-extracted features largely improves detection performance across the board. The main exception is the model using a custom CNN to extract the features. For Replay Attack, the detection performance was slightly reduced when the deep feature extraction was carried out by using the customised CNN model. However, with pre-trained models, the Replay Attack dataset also improved PAD performance. In the evaluation with SiW, both VGG-16 + CLBP and customised CNN + CLBP exhibited hardly any improvement. Nevertheless, ResNet-50 + CLBP and Inception V3 + CLBP improved compared to networks without CBLP. Comparing both Tables 5.3 and 5.4, it can be seen that FPR and FNR were reduced in the proposed method compared to the baseline method. The decreased FPR and FNR resulted in a lower HTER than the baseline in fusion methods, which in turn improved PA detection.

Figure 5.8 shows the ROC curve analysis corresponding to three datasets for baseline and fusion models. In the baseline method (Figure. 5.8 a-c), which uses binary classification using CNN, the best performance was exhibited by the ResNet-50 pre-trained model. The customised CNN model also showed a very close performance to ResNet-50 in the baseline method. However, this CNN model performed better than VGG-16 and Inception-V3 with all three datasets despite having far fewer layers. Among the fusion models, the combination of colour LBP with ResNet-50 features provided the highest detection performance. With CASIA, the customised CNN model features led to performance very close to Inception-V3, but lower than ResNet-50 and VGG-16. However, with Replay Attack and SiW, this model features performed even better than VGG-16 and close to ResNet-50 in combination with CLBP.



(a) CASIA Baseline (b) Replay Attack Baseline (c) SiW Baseline

(d) CASIA Fusion (e) Replay Attack Fusion (f) SiW Fusion

Figure 5.8: ROC curve analysis for CASIA, Replay Attack, and SiW.

Table 5.5 presents a comparison between the computational speed of the baseline and fusion methods for each dataset in training time. Computational speed decreased substantially with VGG-16 and custom CNN model features in the fusion method. Training times of the fusion models reduced to a value of less than 50% of the baseline training time for these two models, with three datasets. At the same time, the PA detection accuracy increased by a value of 6% for CASIA and Replay Attack when VGG-16 features were combined with CLBP features. SiW showed slightly lower accuracy in fusion method with VGG-16 deep features. Even though the custom CNN model-based evaluation scenario had improved computational speed in the fusion method, it did not facilitate improvement in PA detection compared to the corresponding baseline model for Replay Attack and SiW. Only CASIA showed better accuracy when combining the shallow features with the custom CNN. For ResNet-50 and Inception V3, computation speed deteriorated in fusion methods, regardless of the improved accuracy. The times taken for training ResNet-50 and Inception V3-based fusion models were much higher than the corresponding baseline values. However, in ResNet-50-based evaluation scenarios, the largest dataset, SiW, exhibited slightly better accuracy and computational speed in the fusion method than the baseline. The SiW/ResNet-50 combination shows that the fusion method has a slightly faster training time and this implies that, with even larger datasets, the fusion method might have some advantages over the baseline method in both speed and accuracy. However, it should be noted that the fusion feature extraction method has not been fully optimised in these experiments, and it might be possible to improve fusion method training times with some simple software optimisation.

Table 5.5: Computation speed v/s accuracy of baseline and fusion methods

Dataset	**Model**	**Baseline**		**Fusion Method**	
		Computational Speed (s)	**Accuracy (%)**	**Computational Speed (s)**	**Accuracy (%)**
CASIA	VGG-16	**785.86**	85.85	**339.76**	92.33
	ResNet-50	545.66	93.36	746.637	94.65
	Inception V3	229.47	86.74	457.27	88.31
	Custom CNN	2166.59	86.42	408.91	89.34
Replay Attack	VGG-16	**2388.57**	84.25	**945.32**	90.18
	ResNet-50	1611.92	95.57	2261.05	98.56
	Inception V3	663.67	88.78	1366.84	90.38
	Custom CNN	6643.48	94.39	1228.12	93.64
SiW	VGG-16	**16,045.91**	93.02	**5468.71**	92.65
	ResNet-50	14,703.66	98.78	13,382.73	99.60
	Inception V3	5711.11	94.35	7979.04	98.61
	Custom CNN	34,430.00	98.16	7185.88	97.96

5.4.1 Discussion

Presentation attack detection performance exhibited an overall improvement by using the proposed fusion method. Combining local texture features, extracted from different channels with deep features was largely effective in reducing the error in identifying real faces from fake faces. This caused an increment in accuracy and reduction in FPR and FNR.

For CASIA, the fusion method with pre-trained models, reduced FPR, and increased FNR. However, customised CNN, which was trained by using the FAS dataset shows the opposite behaviour with CASIA. FNR deceased and FPR increased. Consequently, features extracted by using pre-trained models as well as the customised CNN model in combination with hand-crafted features provided improvement in PA detection when evaluated with CASIA. A similar performance was given by the Replay Attack dataset. SiW showed a different performance to that of CASIA and Replay Attack because the SiW dataset is “in the wild". Hence, it might not exhibit dataset biases that can be easily exploited by networks trained only on that dataset (i.e., the custom CNN). The other two datasets were recorded under more controlled settings. That is why the results indicated a higher level of capture bias in CASIA and Replay Attack datasets. ResNet-50 and Inception V3 models with colour texture analysis substantially decreased FPR and FNR when evaluated with this dataset. However, for VGG-16 and customised CNN, the FPR and FNR increased slightly, reducing the performance in fusion method.

The custom CNN models used in the experiment were trained and tested by using corresponding FAS datasets. In fact, the features extracted by using this model, when concatenated with CLBP features, in general increased either or both of FPR and FNR. This clearly shows that the model needs further tuning to improve the PA detection performance. The performance analysis also shows that feature extraction by using ResNet-50 is most effective for FPAD among the considered pre-trained models. The fusion models also illustrate the suitability of colour texture analysis in this strategy.

From the ROC analysis curves, it is evident that CNN model trained with the corresponding dataset performed very close to or better than the deep networks considered in the experiments. ResNet-50, Inception-V3, and VGG-16 have 50, 48, and 16 layers, respectively. The customised CNN model has 13 layers, making it shallower than the other models. However, these deep models were trained on the ImageNet dataset, whereas each custom CNN model was trained on the respective FAS dataset training set. This implies that a small dataset and shallower network can achieve comparable or better performance than deep, pre-trained networks.

The computational speed presented in Table 5.5 included the time required to extract deep features and hand-crafted features, and train the classifier model. For each dataset, the hand-crafted feature extraction time is the same. Moreover, the classifier training period is significantly less than the time taken for feature extraction. The variation in the recorded computational speeds relies upon deep feature extraction speed. Hence, a pre-trained model with depth equal to or less than VGG-16 could be used to extract deep features to improve the performance of this fusion method by using CLBP. From the results, it is also evident that the performance of shallow models trained on FAS datasets was not improved by fusion. This implies that the features that emerge in these shallow models may already encompass the shallow, engineered features. Other suitable hand-crafted features could also be combined with the custom CNN models trained on FAS datasets to investigate the impact of custom CNN model deep features. A challenge is extracting suitable handcrafted features which can provide spoof-specific patterns and further increase FPAD performance, specifically in unseen attack detection.

The comparative analysis using accuracy, HTER, and computational speed which are presented in Tables 5.3–5.5 points to the advantages, drawbacks, and challenges of the proposed fusion method. The fusion method performed better in PA detection when deep features were extracted by using pre-trained models than the models which were trained with FAS datasets. Among the pre-trained models considered for evaluation, the model with the fewest layers (VGG-16) showed improvement in computational speed and detection performance. One possible hypothesis for this is that FPAD relies on low-level, spoof-specific features, rather than complex deep features. Deeper pre-trained models were able to improve detection performance at the cost of computational speed. However, for application in real-life scenarios, the best model would exhibit optimal performance both in accuracy and computational speed.

5.5 Conclusion

An experimental framework combining colour texture and deep features is presented in this chapter. Colour texture features when combined with deep features, substantially reduce the number of false positives in most cases. This suggests that rather than global features, task-specific features are more likely to facilitate the detection of PAs. Moreover, a fusion method using pre-trained deep models also improves computation speed on some models and shows promise for experiments with larger datasets. By incorporating additional features corresponding to texture, frequency, and image quality, this experimental framework could be extended to detect PAs. A cross-dataset analysis will confirm whether these relatively shallow features are generalisable.

Chapter 6

Unmasking the Imposters: Task-specific feature learning

The previous chapter provides evidence that the utilisation of task-specific features, rather than global features extracted from pre-trained models, is essential for improving the detection of presentation attacks (PA). Building upon this insight, the current chapter introduces an innovative approach to learning task-specific features by leveraging deep pre-trained models. The primary objective of this approach is to enhance the generalisation capabilities in FPAD. Specifically, the method involves fine-tuning the higher convolutional layers of pre-trained models for task-specific feature learning. Through extensive experiments, in comparison to transfer learning and hybrid models, this technique demonstrates improved performance across different datasets, highlighting enhanced cross-dataset performance and greater generalisability.

The main findings of this chapter are accepted to be presented at the 2023 International Joint Conference on Neural Networks (IJCNN) as part of the paper "Unmasking the Imposters: Task-specific feature learning for face presentation attack detection [232]".

6.1 Overview

The FPAD detects PA based on differences in features between fake and genuine facial images. In order to accomplish this, earlier FPAD models used handcrafted features [197] related to texture, image quality, motion, and frequency. The extracted features were classified with SVM, RF or K-NN classifiers. These hand-crafted features are domain-specific features [233]. Hence, hand-crafted feature methods had limited generalisation as they use only domain-specific features rather than task-specific features, especially in the RGB domain.

The automatic feature extraction capability of deep learning models further enhanced FPAD performance. In a deep model, lower layers provide domain-specific features such as edges, and corners. However, the higher layers learn task-specific features. The FPAD task is to differentiate between real and fake facial images of the same user. But image classification categorises different objects in the given images. Thus, task-specific features of FPAD are different from that of image classification. Hence, learning task-specific features is more important in improving generalisation in FPAD.

As a deep learning technique, transfer learning has been exploited in a number of ways to address FPAD by learning either domain-specific features or task-specific features. For transfer learning, existing FPAD models have used pre-trained image classification or face recognition models. Since task-specific features are provided by higher layers, image classification models were used after modifying the top fully connected layers and fine-tuning them to detect PA [75]. Domain-specific features were learned [17] by fine-tuning a few lower convolutional layers in a pre-trained face recognition model using multi-spectral data. Nonetheless, the majority of Face Anti-Spoofing (FAS) datasets are in the RGB domain. So, it may be more effective to use a model that can extract task-specific features from RGB datasets rather than using multi-modal data. The research in chapter 5 [12] has shown that fusion models using deep pre-trained models and hand-crafted methods improved PA detection in intra-dataset evaluations. Thus transfer learning has been explored extensively in face anti-spoofing.

This chapter presents a transfer learning model, to learn task-specific features to improve generalisation. The higher convolutional layers of deep pre-trained models were fine-tuned along with the fully connected layers using a public FAS dataset SiW. This fine-tuned model was used to extract features, which were used to form fusion models. Fusion models were formed using the deep features from fine-tuned models and combining the deep features with hand-crafted features. The experiments used the public FAS datasets, CASIA, and Replay Attack, for cross-dataset validation.

6.2 Task-specific learning and fusion for FPAD

Figure 6.3 provides a schematic diagram of the fusion model using task-specific feature learning. Typically, this includes utilising fine-tuning pre-trained models, hand-crafted features extraction and features fusion to form models to detect presentation attacks. To improve generalisability, deep pre-trained models were fine-tuned in order to learn task-specific features. Fusion models were also formed using features extracted from fine-tuned models and handcrafted features. Accordingly, fine-tuned and fusion models were evaluated for intra-dataset and cross-dataset performance as shown in Figure. 6.3. By using the SiW train set, the models were fine-tuned. The SiW, CASIA, and Replay Attack test sets were used to evaluate the performance.

The models were evaluated using public FAS datasets, CASIA [41], Replay Attack [46], and SiW [102]. These datasets consist of 2D PA variants including print, photo and video attacks. Figure 5.5 shows samples of real and fake faces derived from three datasets. Figure 6.1 shows genuine facial images in the top row. In the lower row, corresponding fake facial images are displayed. A comparison of the three datasets is presented in Table.6.1. The upper row in each figure contains the real-face samples, whereas the lower row has the PA samples.

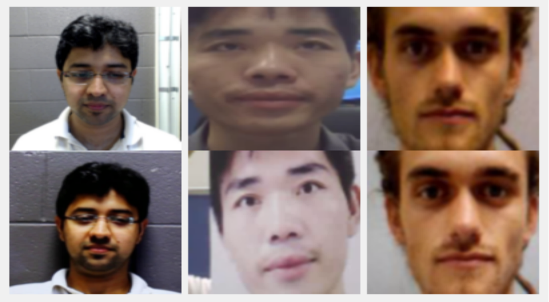


Figure 6.1: Real and PA image samples from SiW, CASIA and Replay Attack

Table 6.1: Comparison of FAS datasets used in the evaluation

Dataset	**CASIA**	**Replay Attack**	**SiW**
Subject	50	50	165
Live videos	150	200	1320
Attack videos	450	1000	3300
Attack types	2 Print, Replay	Print, 2 Replay	2 print, 4 Replay
Display devices	iPad	iPhone 3GS, iPad	iPad Pro, iPhone 7, Galaxy S8, Asus MB168B

6.2.1 Fine-tuning

Existing FPAD methods used either domain-specific or task-specific features through fine-tuning deep pre-trained models in different ways. Since lower layers provide domain-specific features, some recent research followed the concept of domain-specific adaptation using multi-spectral data and a pre-trained face recognition model.

On the other hand, task-specific features from RGB data were extracted by modifying and fine-tuning the classifier layers of deep pre-trained classification models. These methods showed reduced cross-dataset performance, while domain-specific adaptation required multi-spectral data. To circumvent both limitations, higher convolutional layers of the deep pre-trained classification model were fine-tuned using the SiW train set. VGG-16 and Inception V3 were fine-tuned in a similar way.

More specifically, the fine-tuned VGG-16 and ResNet-50 models had six higher convolutional layers re-trained. The fine-tuned Inception V3 model had eight higher convolutional layers which were retrained using the SiW dataset. The top layers included layers as follows: a fully connected layer of size 4096, batch normalization layer, dropout layer, another fully connected layer of size 4096, batch normalization layer, dropout layer, a fully connected layer of size 512, another fully connected layer of size 256 and a sigmoid layer.

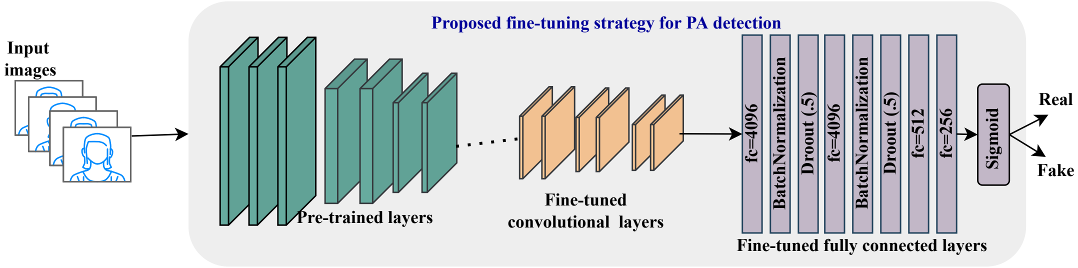


Figure 6.2: Fine-tuning

6.2.2 Fusion

Fine-tuned ResNet-50 model performance was compared with fusion model (Fig. 6.3) performance. Fusion models were formed in three ways;

- Fusion of fine-tuned ResNet-50 feature and hand-crafted features.
- Fusion of fine-tuned ResNet-50 and VGG-16 features.
- Fusion of fine-tuned ResNet-50, VGG-16 and Inception V3 features.

The features were combined using concatenation. Hand-crafted features included colour texture (CLBP), Difference of Gaussian (DoG), Histogram of Oriented Gradients (HOG) and Fast Fourier Transform (FFT). PAs introduce chrominance disparities while preserving luminance variations. Hence, the chrominance disparities cannot be identified in RGB colour space. FAS needs alternative colour spaces such as HSV and YCbCr to utilize chrominance disparities invisible in RGB colour space. HSV and YCbCr have chrominance components. As HSV and YCbCr colour spaces contain spoof-specific chrominance disparities, the images can be converted into these colour spaces and texture analysis can be performed to detect PAs [8]. The HOG provides information about the structure of the objects in the image. HOG provides edge features as well as edge direction. By extracting the edge orientation and gradients, this edge direction is provided. Thus, HOG features derived from an image represent local disparities in gradient and orientation, which can be applied to detect PAs [48]. Recapturing eradicates high-frequency features from the images, creating a disparity between real and fake facial images. These disparities can be used to identify PAs [51]. Edge detection has been used to identify differences in local features to detect PAs. DoG is applied to an image to mitigate noise and preserve high-frequency features, especially edges. Being an edge detection filter, DoG enhances the edges in the final image. The deformities in the PAs introduce differences in local features compared to the real facial image. Hence, edge detection has been used to detect recaptured images and PAs. Frequency disparities between real and fake facial images can also be extracted using FFT [234].

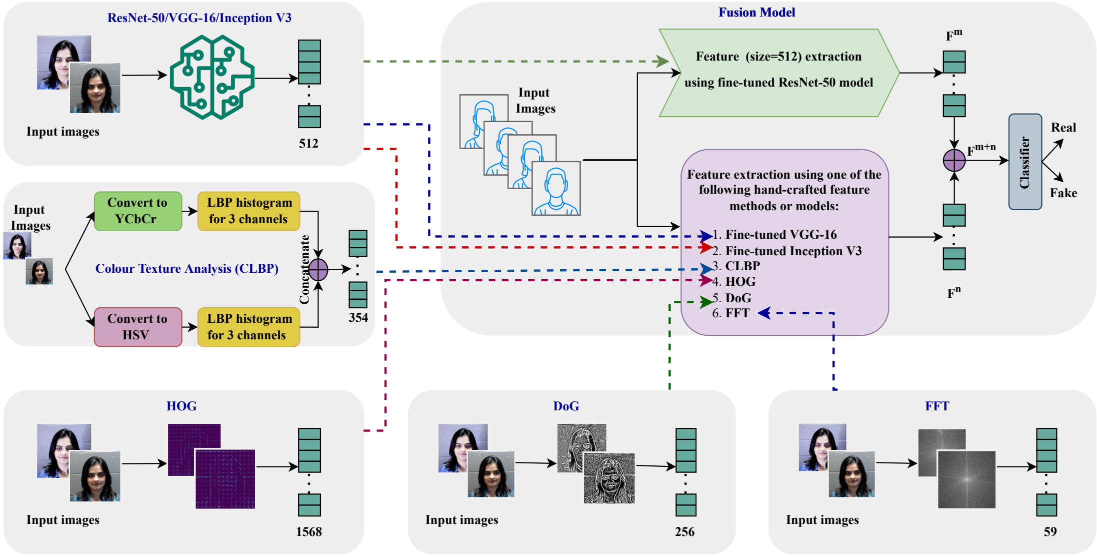

ResNet-50/VGG-16/Inception V3
Input images
512
Input Images
Convert to YCbCr
LBP histogram for 3 channels
Colour Texture Analysis (CLBP)
Convert to HSV
LBP histogram for 3 channels
Concatenate
354
Fusion Model
Input Images
Feature (size=512) extraction using fine-tuned ResNet-50 model
Feature extraction using one of the following hand-crafted feature methods or models:
1. Fine-tuned VGG-16
2. Fine-tuned Inception V3
3. CLBP
4. HOG
5. DoG
6. FFT
F^m
F^{m+n}
F^n
Classifier
Real
Fake
HOG
Input images
1568
DoG
Input images
256
FFT
Input images
59

Figure 6.3: Fusion model

Let F^m be a feature vector with size m and F^n be another texture feature vector with size n. Thus F^{Fusion} will have the size $(m+n)$. Then the final feature vector F^{Fusion} can be represented [16] as in Equation. 6.1.

$$F^{Fusion} = F^m \cup F^n \tag{6.1}$$

Following the above equation, the fusion vectors in three experimental scenarios, F^1, F^2 and F^3 can be represented as in Equation. 6.2,Equation. 6.3, and Equation. 6.4.

$$F^1 = F^{ResNet-50} \cup F^{hand-crafted} \tag{6.2}$$

$$F^2 = F^{ResNet-50} \cup F^{VGG-16} \tag{6.3}$$

$$F^3 = F^{ResNet-50} \cup F^{VGG-16} \cup F^{InceptionV3} \tag{6.4}$$

F^1 represents the feature vector corresponding to the fusion models combining fine-tuned ResNet-50 features and hand-crafted features. F^2 is the combined feature vector of fine-tuned ResNet-50 features and fine-tuned VGG-16 features. Fine-tuned ResNet-50, VGG-16 and Inception V3 feature vectors were combined to form F^3. Deep feature vectors from ResNet-50, VGG-16 and Inception V3 had a size of 512. hand-crafted feature vectors had varying feature sizes. The resultant feature vector size will be the sum of the feature vectors used in the experiments.

The classifier module for all models included 9 layers including four fully connected layers two batch normalization, two dropout layers and an output sigmoid layer as in Figure. 6.2. A detailed structure of the classifier is as follows: a fully connected layer of size 4096 followed by batch normalization and dropout, another fully connected layer of size 4096 followed by batch normalization and dropout, a fully connected layer with size 512, another fully connected layer with size 256 and an output sigmoid layer.

Colour Texture Analysis (CLBP)

Colour texture analysis [8] was one of the hand-crafted features used in the fusion models. The human eye is more sensitive to luminance than chrominance. Hence, luminance variation in the source image is preserved in PAs that are printed or displayed in RGB. However, PAs include chrominance disparities, which can be used to identify fake from authentic facial images. On the other hand, chrominance variations are invisible to the human eye. In the RGB colour space, there exists a high correlation between colour components. The recapturing process involved in PAs introduces chrominance disparities. At the same time, luminance variations will be preserved. Hence, chrominance disparities cannot be identified in the RGB colour space. It might be possible for a deep CNN to learn them eventually, but it will be easier in an appropriate data representation. FAS needs alternative colour spaces such as HSV and

YCbCr to utilize the chrominance disparities invisible in the RGB colour space. Both HSV and YCbCr have chrominance components. The chrominance component in HSV is complementary to the chrominance component in YCbCr. As HSV and YCbCr colour spaces provide spoof-specific chrominance disparities, the images can be converted into these colour spaces and carry out texture analysis to identify PAs.

To conduct colour texture analysis (CLBP) (Fig. 6.4), RGB images were converted into HSV and YCbCr colour spaces and then the LBP of each channel in these images were extracted. LBP histograms from these six channels were combined to form a final feature vector of size 354.

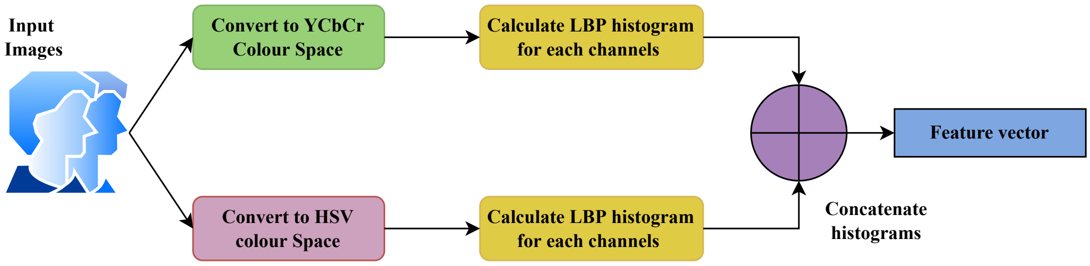
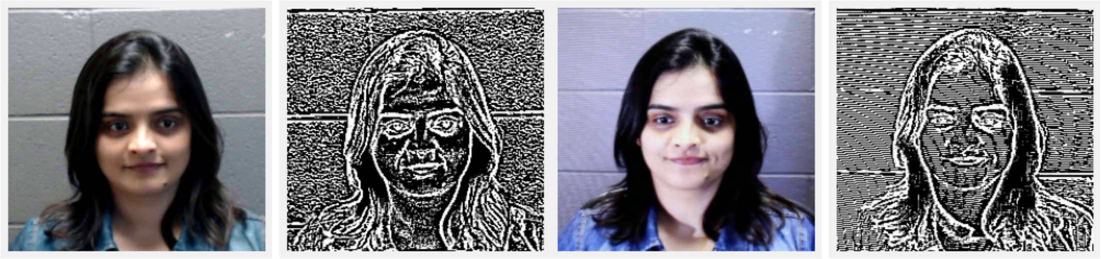

Input Images
Convert to YCbCr Colour Space
Calculate LBP histogram for each channels
Convert to HSV colour Space
Calculate LBP histogram for each channels
Feature vector
Concatenate histograms

Figure 6.4: Colour Texture Analysis

Difference of Gaussian (DoG)

Difference of Gaussian (DoG) is a technique used to detect the edges in images. In DoG, a blurred version of the source image is created by performing Gaussian blur with two different standard deviation values. The difference between these blurred images is taken as the final image. Being an edge detection filter, DoG enhances the edges in the final image. PAs mainly include recaptured images. The deformities in the PAs introduce differences in local features compared to the real facial image. Hence, edge detection has been used in detecting both recaptured images and PAs. DoG is applied to an image to mitigate the noise component and preserve the high-frequency features, especially edges. Recapturing eradicates high-frequency features from the images, creating a disparity between real and fake facial images. These disparities can be used to identify PAs [51].



(a) Real image (b) DoG representation of real image (c) Fake image (d) DoG representation of fake image

Figure 6.5: DoG images of real and fake facial images

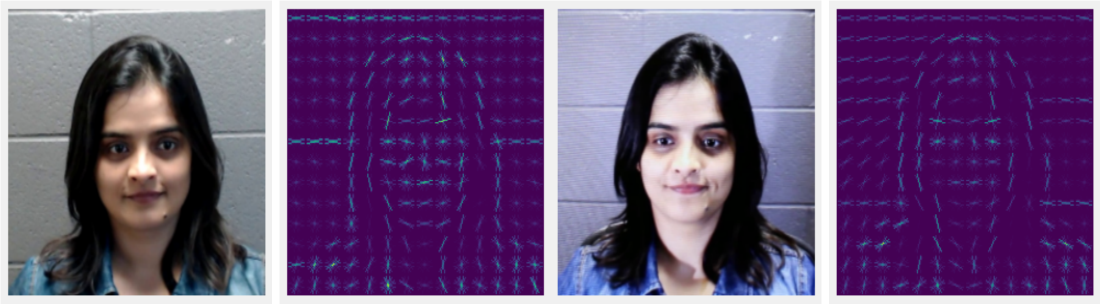
DoG acts as a band pass filter. This band-pass filter incorporates two Gaussian filters with standard deviations as limits. For, the inner Gaussian filter, standard deviation, σ_i was set as .6 and for the outer Gaussian filter standard deviation, σ_j was set as 1. The Gaussian standard deviation values were selected in such a way that, the filtering reduces low spatial frequency information and preserves high-mid frequency information. This facilitates identifying PAs using these high-frequency cues. For an image *i(x,y)*, DoG image with standard deviations σ_i and σ_j can be defined as in Equation. 6.5 [9];

$$I(x, y, \sigma_i, \sigma_j) = (G(x, y, \sigma_i) - G(x, y, \sigma_j)) * I(x, y) \tag{6.5}$$

$G(x, y, \sigma_i)$ is the Gaussian filter with standard deviation, σ_i and $G(x, y, \sigma_j)$ is the Gaussian filter with standard deviation, σ_j. In these experiments, DOG resultant images were converted to grey-scale images. Then, the histogram was extracted to get the feature vector from the images.

Histogram of Oriented Gradients (HOG)

Histogram of Oriented Gradients (HOG) provides the structure of the objects in the image. In addition to edge features, HOG gives the edge direction. This edge direction is provided by extracting the orientation and gradients of the edges. The images are divided into smaller regions. Orientation and gradients are calculated locally for each of these regions. A histogram for each region is generated. Hence, HOG features extracted from an image represent local disparities in terms of gradient and orientation.



(a) Real image (b) HOG of real image (c) Fake image (d) HOG of fake image

Figure 6.6: HOG images of real and fake facial images

Local features show disparities in PAs compared to genuine facial images. So, HOG has been used in the existing literature to detect PAs [48]. Therefore, HOG was considered one of the hand-crafted features used in fusion models. The experiments used a region cell size of 16 × 16. Histograms from these cells together form a final feature vector of size 1568.

Fast Fourier Transform (FFT)

Fourier Transform produces the frequency domain representation of an image, which is in the spatial domain. In the Fourier domain image, each point represents a particular frequency contained in the spatial domain image. The frequency domain transformation of a digital image is carried out through a 2-dimensional Discrete Fourier Transform (2D DFT) using Fast Fourier Transform (FFT) Algorithm. The spatial domain image of a transformed image provides the frequencies present in the image. Since DFT is a complex image, such images were analysed using real and complex parts or phase and magnitude responses. The magnitude response is mostly used for analysis as it preserves the majority of the spatial information in the spatial domain. Thus it provides information, which is not visible in the spatial domain.

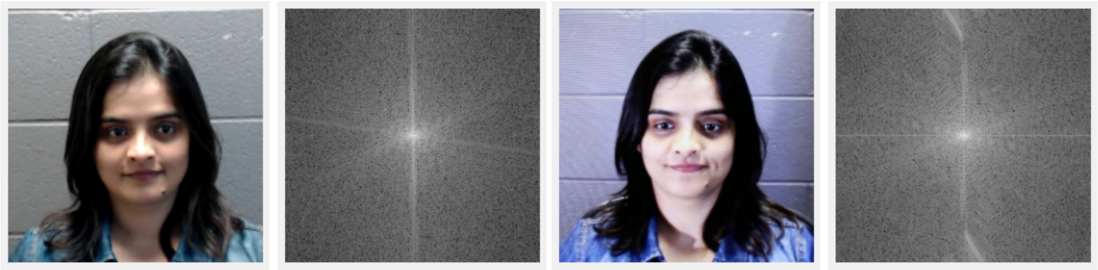
For an image *i(x,y)* of size (N,N) in the spatial domain, the Fourier Transform, I(u,v) can be represented as,

$$I(u,v) = \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} i(x,y) e^{-i2\pi(\frac{ux}{N}+\frac{cy}{N})} \tag{6.6}$$

I(u,V) has real and imaginary part. The magnitude is usually used to represent *I(u,v)*, where |*I(u,v)*| is,

$$|I(u,v)| = \sqrt{Real(I(u,v))^2 + Imaginary(I(u,v))^2} \qquad (6.7)$$

For these experiments, the RGB image was converted into a gray scale and the calculated magnitude of the transform function. Since the spatial domain image of the magnitude spectrum represents the frequency features, the LBP histogram of this image was used to extract the feature vector.



(a) Real image

(b) FFT magnitude spectrum representation of real image

(c) Fake image

(d) FFT magnitude spectrum representation of fake image

Figure 6.7: Frequency domain representation of real and fake facial images

6.3 Experimental settings

Experiments included fine-tuning different numbers of layers in the pre-trained ResNet-50 model, fine-tuning higher convolutional layers and top fully connected layers of the pre-trained VGG-16 and Inception V3 model, a fusion of hand-crafted features, HOG, CLBP, DoG and FFT with fine-tuned ResNet-50 features, and fusion of fine-tuned VGG-16 and fine-tuned Inception V3 features with fine-tuned ResNet-50 features. FAS dataset SiW was used for training. For, testing, SiW, CASIA and Replay Attack were used. CASIA, Replay Attack and SiW are widely used public FAS datasets which have photo and video attacks in them. The pre-trained models were fine-tuned keeping ImageNet weights as initial weights.

From the dataset videos, faces were detected from CASIA and Replay Attack frames were extracted at a rate of 2 frames per second. Using the SiW dataset, frames were extracted at 1 frame per second and face detection was performed based on the annotations provided. A random scaling of the bounding box for SiW was also performed to provide some background information and improve the diversity of facial images. The facial images from three datasets were resized to 224 × 224 pixels. The official train-test split was maintained for all three datasets. Table 6.2 summarizes the number of training and test images in each dataset.

Table 6.2: Dataset sample size in train and test partitions .

Dataset	**Train**			**Test**		
	Real	Fake	Total	Real	Fake	Total
CASIA	527	1760	2287	824	2471	3295
Replay Attack	1689	5261	6950	1928	5645	7573
SiW	14733	26057	40790	12390	22389	34779

The experiments included fine-tuning different of layers the pre-trained ResNet-50 model. All the models in the experiments used the SiW train set for training. For intra-dataset evaluation, the models were tested using the SiW test set. CASIA and Replay Attack test sets were used for cross-dataset evaluation. Fine-tuning was carried out as explained in Section. 6.2. The fine-tuned models were utilized later for deep feature extraction for the fusion method. Binary cross entropy loss and Adam optimizer were used for model compilation. For fine-tuning and fusion models, the learning rate used was 5×10^{-6} for all datasets. The batch size and epochs were 512 and 10 respectively. The results are reported using accuracy, Average Classification Error Rate (*ACER*) (Section. 2.10) and ROC curve analysis.

6.4 Task-specific learning: A generalisation analysis

Intra-dataset and cross-dataset comparisons of the fine-tuned and fusion models are presented in Table. 6.3 and Table. 6.4 respectively. The results were reported in terms of accuracy, AUC and ACER. In the table, ResNet-50 (FC)indicates, the transfer learning model using the pre-trained ResNet-50 model and ResNet-50 (ALL) is the model which had all the layers fine-tuned using SiW train set. ResNet-50, VGG-16 and Inception V3 represent the models with fine-tuned higher convolutional as well as modified fully connected layers.

Table 6.3: Intra-dataset performance using SiW dataset

Models	**ACC (%)**	**ACER(%)**	**AUC**
ResNet-50 (FC)	97.53	2.33	0.99
ResNet-50	99.14	1.06	1.00
ResNet-50(ALL)	**99.57**	**0.51**	1.00
ResNet-50+CLBP	99.28	0.86	1.00
ResNet-50+HOG	99.27	0.90	0.99
ResNet-50+DoG	99.28	0.87	1.00
ResNet-50+FFT	99.28	0.89	0.99
ResNet-50+VGG-16	99.51	0.64	1.00
ResNet-50+Inception V3	99.23	1.01	0.99
ResNet-50 +VGG-16+ Inception V3	99.53	0.60	1.00

From the Table. 6.3, it is evident that the intra-dataset accuracy (99.57%) and ACER (.51%) showed as the best detection performance when all the layers of the pre-trained ResNet-50 model were fine-tuned using SiW train set. However, ResNet-50 (ALL) exhibited lower cross-dataset performance when tested with CASIA and Replay Attack, compared to fine-tuned models (ResNet-50 (FC) and ResNet-50) and fusion models (Table. 6.4). In the cross-dataset evaluation of CASIA, ResNet-50 showed the best performance. The model accuracy when tested with CASIA was 88.80%. The ResNet-50 model exhibited ACER of 13.98%. Nevertheless, the ResNet-50 model showed an accuracy of 85.05% and ACER of 24.61% when tested with Replay Attack.

The fusion model combining ResNet-50 and VGG-16 deep features exhibited the best cross-dataset performance (accuracy:87.43% and ACER:20.11%) with Replay Attack. Except with (ResNet-50 (FC) and ResNet-50 (ALL), cross-dataset evaluation with CASIA and Replay Attack provided accuracy greater than 80% which shows better generalisation. Fusion models slightly reduced cross-dataset performance when tested with CASIA. However, compared to ResNet-50 models, fusion models using only deep features showed an increase in performance when tested with Replay Attack. Cross-dataset performance with Replay Attack also decreased slightly with fusion models using hand-crafted features and ResNet-50 features.

Table 6.4: Cross-dataset performance with CASIA and Replay Attack.

Models	**CASIA**			**Replay Attack**		
	ACC (%)	ACER(%)	AUC	ACC (%)	ACER(%)	AUC
ResNet-50 (FC)	75.45	48.89	0.43	74.86	48.78	0.65
ResNet-50	**88.80**	**13.98**	0.93	85.05	24.61	0.82
ResNet-50(ALL)	76.21	43.35	0.62	73.52	50.58	0.57
ResNet-50+CLBP	86.94	16.26	0.93	82.32	30.25	0.79
ResNet-50+HOG	86.65	15.47	0.91	84.19	26.62	0.77
ResNet-50+DoG	87.73	15.05	0.93	82.99	28.82	0.69
ResNet-50+FFT	87.34	16.68	0.88	82.83	29.14	0.77
ResNet-50+VGG-16	85.54	15.71	0.92	**87.43**	**20.11**	0.82
ResNet-50+Inception V3	87.39	14.20	**0.94**	85.92	23.05	0.75
ResNet-50 +VGG-16+ Inception V3	87.00	14.21	0.92	85.60	22.80	**0.85**

ROC comparison of fine-tuned ResNet-50 models is shown in Fig. 6.9. The ROC analysis indicates that in intra-dataset evaluation with SiW, the models correctly detect PAs. Among the models evaluated cross-dataset, ResNet-50 with fine-tuned higher convolutional layers and fully connected layers (ResNet-50) demonstrated the highest performance. The fusion models were compared with the ResNet-50 model (Fig. 6.10). Compared to the ResNet-50 model, the fusion models have very similar performance, both in intra-dataset and cross-dataset evaluations. A fusion model based on the deep features of ResNet-50 and VGG-16 performed better. Fusion models, however, when formed using ResNet-50 features and hand-crafted features, showed reduced performance for cross-dataset performance in comparison to ResNet-50.

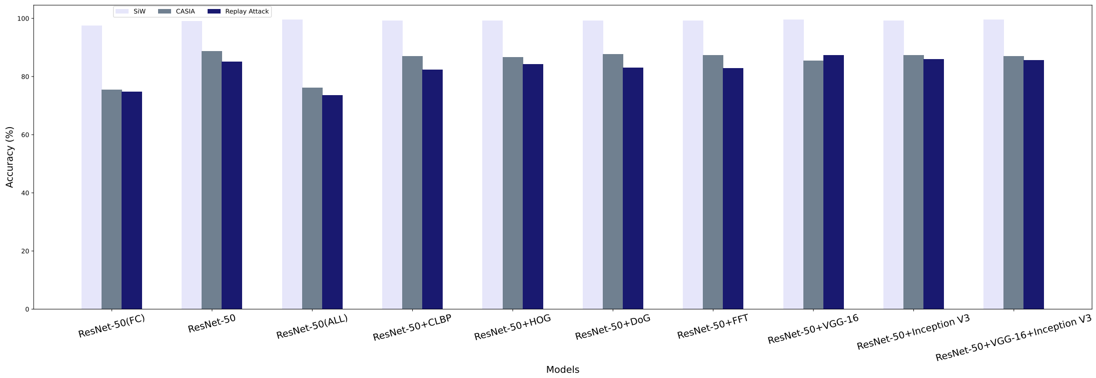
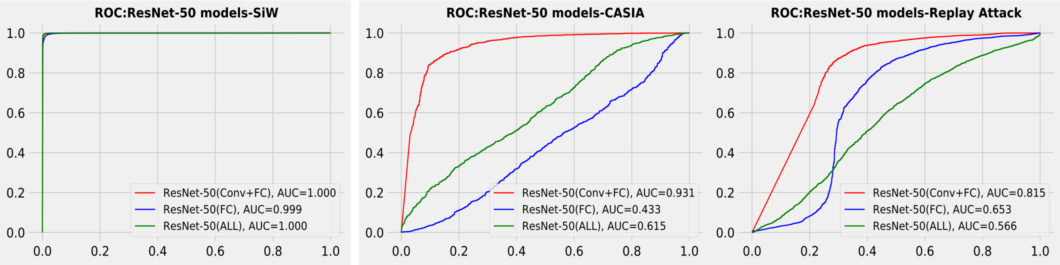
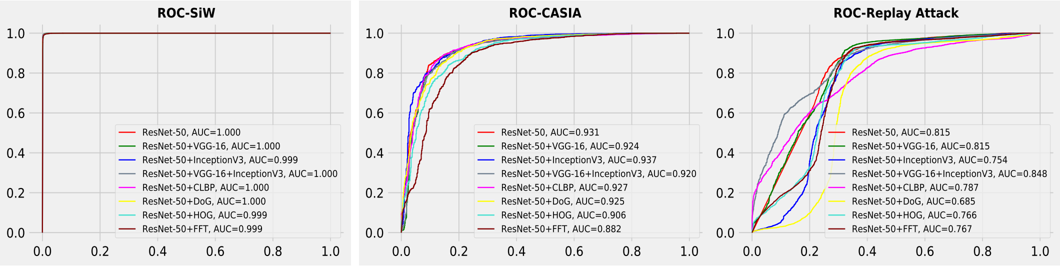


Figure 6.8: Performance comparison of fine-tuned and fusion models



(a) SiW (b) CASIA (c) Replay Attack

Figure 6.9: ROC comparison of fine-tuned ResNet-50 models



(a) SiW (b) CASIA (c) Replay Attack

Figure 6.10: ROC comparison of Task-specific feature learning and fusion models

6.5 Discussion

PA detection evaluates the genuineness of the facial image captured by the sensor. Therefore, the PA detection task examines the disparities between fake and real facial images in terms of features such as texture, image quality and frequency in hand-crafted feature methods. In a deep learning model, lower convolutional layers learn domain-specific features and higher layers learn task-specific features. However, pre-trained image classification models cannot fully provide the features required to detect spoofing in RGB domain. The major cause is that those models were trained to detect the object in the images using the overall image features rather than checking the genuineness of the images. Hence, various pre-trained models were used after fine-tuning using FAS datasets for the PA detection task.

The experiments used the pre-trained ResNet-50 model, which was fine-tuned using the FAS dataset, SiW. Fine-tuning was carried out with three different methods to analyse the performance (Section. 6.2). Among the three methods used to fine-tune the pre-trained ResNet-50 model, the best was fine-tuning the higher convolutional layers and fully connected layers, which provided impressive performance in intra-dataset and cross-dataset evaluations as in Table. 6.3 and Table. 6.4.

Table 6.5: Comparison of task-specific learning with SOTA methods

Model	**ACER(%)**
FAS-TD-SF [105]	39.4
LGON [235]	20.56
Fusion model	14.20
ResNet-50 (Ours)	**13.98**

Domain-Specific Units (DSU) [17] have been used to achieve domain adaptation in PA detention fine-tuning FR models using multi-modal data. Nonetheless, higher convolutional layers provide task-specific features. Therefore a pre-trained ResNet-50 model was used, after fine-tuning its higher convolutional layers, and fully connected layers using only the RGB FAS dataset to extract task-specific generalisable features. This fine-tuned model exhibited performance comparable to the SOTA methods in cross-dataset evaluation with CASIA as shown in Table. 6.5. The fine-tuned model was also compared with fusion models, where extracted deep features from this fine-tuned model were combined with either hand-crafted features or deep features, extracted from fine-tuned VGG-16 and Inception V3 models.

Task-specific features should be learned for better-unseen attack detection in FAS. It is a known fact that higher convolutional layers provide task-specific features. Hence, fine-tuning higher convolutional layers can enable the extraction of task-specific features which are essential for spoof detection tasks to attain generalisation. In Table. 6.5, two SOTA methods are compared with the proposed fine-tuned and fusion models. The considered methods used similar train-test dataset combinations in cross-dataset evaluation. It is evident from the cross-dataset performance ACER values that fine-tuned ResNet-50 performs better compared to FAS-TD-SF [105] and LGON [235]. However, both fine-tuned and fusion models perform slightly lower than LGON model in cross-dataset evaluation with Replay Attack.

Both CASIA and Replay Attack datasets performed better with models using deep fine-tuned features and their fusion rather than fusion models with hand-crafted features, showing that models using deep fine-tuned features are more effective in PA detection and generalisation. It also helps to avoid the disadvantages of using hand-crafted features and their extraction. Fine-tuning higher convolutional and fully connected layers of the deep pre-trained models using FAS data for FAS increases the generalisation using the inherent feature extraction capability of the deep CNN model.

6.6 Conclusion

The experiment framework presented in this chapter compares the intra-dataset and cross-dataset performance of fine-tuned ResNet-50 model, and fusion models. The pre-trained models were used for PA detection and feature extraction after fine-tuning the higher convolutional layers and fully connected layers using the FAS dataset, SiW. The results illustrate that fine-tuning higher convolutional layers provide task-specific features, which in turn improves generalisation in FPAD compared to fusion with hand-crafted features and fine-tuning only fully connected layers.

Chapter 7

Conclusion

This chapter summarises the findings of the preceding chapters and discusses the limitations and future works evident within the work.

The thesis focuses on the detection of face presentation attacks (FPAD) using deep learning. Specifically, data aggregation, hybrid fusion, and task-specific learning methods were developed to improve cross-dataset performance for detecting presentation attacks. A major objective of these methods is to improve the feature space so that generalization can be improved by increasing variation in the training data, combining hybrid features, and fine-tuning pre-trained models to learn task-specific features. Furthermore, this chapter provides a brief overview of the limitations of the proposed methods and some potential future research in this area.

7.1 Summary

In line with the objectives of Chapter 1, the following section provides a summary of contributions and key findings:

This thesis reviews existing state-of-the-art methods for face presentation attacks in chapter 2. These methods include hand-crafted feature methods, transfer learning, multi-modal methods, anomaly detection, and hybrid methods. Additionally, this chapter discusses the existing public datasets as well as their modalities, attack variants, and other domain-dependent factors such as devices, settings, and spoof mediums. In addition, the evaluation metrics used in FPAD were explained. Besides addressing the challenges associated with face presentation attack detection, future research directions were also discussed.

Chapter 3 demonstrates that using custom dataset partitions to train face presentation attack detection models impacts results significantly. Custom dataset partitions on the NUAA dataset increase the variance of the training set, providing adequate samples to data-hungry models. NUAA training and testing sets, on the other hand, present challenges to generalization due to their disjointed distributions. Combining the two distributions results in custom partitions, minimizing generalisation challenges. Moreover, increasing the training set variance improves detection performance.

To detect presentation attacks effectively, data aggregation and deep transfer learning were applied in chapter 4. Three widely used face anti-spoofing public datasets – NUAA, CASIA, and Replay Attack – were combined to form a new aggregated dataset. Using four public datasets, both intra-dataset and cross-dataset evaluations were conducted. Official partitions of each dataset were taken into account when forming aggregated train and test sets. Multiple source domains alone are not sufficient to guarantee domain generalisation against unknown attacks, as demonstrated by the experiments. As FPAD is intended to be generalised, a method must be designed so that generalisable features can be extracted.

Chapter 5 presents a hybrid fusion of colour texture features and deep features to detect presentation attacks. The experiments showed that Colour texture features when combined with deep features, substantially reduce the number of false positives in most cases. This suggests that rather than deep features, features specific to spoofing tasks are more likely to facilitate the detection of PAs.

Detection of PAs is based on the learning of spoof-specific features from attacks and utilizing these features to identify them. Chapter. 6 illustrates a method of learning task-specific features through transfer learning for enhancing PA detection and thereby generalization. A pre-trained model is shown to be capable of learning task-specific features when fine-tuned by using higher convolutional layers.

7.2 Limitations and Future Work

The techniques established within this thesis make a valuable contribution towards enhancing presentation attack detection and generalization. The models' effectiveness was assessed by employing three publicly available datasets. Nevertheless, there are areas that require further improvement, and here are some potential avenues for future research.

7.2.1 Limitations

This thesis used public FAS datasets based on 2D attacks to develop the research, but not datasets based on 3D attacks. As a consequence, it has yet to be investigated how FPAD performs under 3D attacks as well. A current trend in evaluation is to combine multiple datasets for training and test the model with entirely different datasets in order to assess the generalisation of the model to different domains. Although a similar data aggregation strategy is examined in this thesis using transfer learning in Chapter 4, most of the other methods presented in this thesis have not been evaluated in the same way. The NUAA and CASIA datasets, both of which were used in this study, do not have variances in ethnicity within them. The rest of the datasets used in this study have a limited degree of variance, with the exception of SiW. In addition, only a limited number of pre-trained models as well as handcrafted features were used in this study. The concept of fusion was explored only from the perspective of feature fusion.

Latency serves as an additional metric for comparing models. In Chapter 5, latency was compared among various models trained on diverse datasets. This comparison of latency is crucial for assessing how well the models perform in real-time deployment scenarios. However, similar latency comparisons have not been conducted for the other methods. Additionally, when presenting model performance, performing statistical analysis on the metric values is another method to demonstrate the consistency of detection performance. Nevertheless, this research reports model performance solely as the average values of evaluation metrics, without incorporating statistical analysis.

7.2.2 Future Work

The following are some examples of future work that could be carried out in order to take this research in different directions:

Measuring the Effectiveness of Task-Specific Learning in Domain Generalisation

A possible extension of the task-specific feature learning presented in this thesis is the combination of data aggregation and task-specific learning. As part of the thesis, three public datasets were considered and pre-trained models were fine-tuned using each dataset. It is, however, possible to learn features from other attack variants by using even more diverse datasets. Therefore, it is necessary to evaluate the domain generalisation effectiveness of this method. This testing involves training the model on three or more source databases and subsequently evaluating its performance on a completely unseen database using the leave-one-out (LOO) strategy [235, 140]. For the task-specific learning method, the binary cross-entropy loss was used as the loss function. A custom loss, however, would be another extension of this work.

Data Partitioning: Uncovering Advanced Strategies

Partitioning a dataset is a crucial task in data-driven models. One of the most straightforward and commonly used methods to divide such a dataset is by randomly sampling a portion of it. In this thesis, the research delved into the significance of dataset partitioning within the context of FPAD. The study involved both an official dataset partition and a custom partition, where a specific percentage was randomly assigned to the training and testing sets. Since the training set plays a pivotal role in discriminative feature learning and model performance, it becomes imperative to select the optimal training set to enhance both performance and efficiency, all while avoiding overfitting and bias. Given these considerations, the exploration of optimizing dataset partitioning [236, 237] becomes a valuable pursuit. Implementing online optimization for data partitioning [238], represents a viable strategy in this scenario.

Making the Most of Multi-Modal Data

NIR, thermal, and depth imaging provide multiple cues that can be used to identify PAs and enhance their detection. In this scenario, however, there is a requirement for additional hardware to be integrated with the FR system. As a consequence, mobile devices are limited to using RGB-based FPAD models due to the lack of multi-modal sensors. In recent years, mobile devices have become equipped with LiDAR sensors. The authors of [182] have developed a multi-modal dataset that incorporates LiDAR images. Among the other multi-modal FAS datasets is Echoface-Spoof [146], which contains acoustic data recorded using inbuilt acoustic sensors of mobile devices. These novel multi-modal datasets offer a promising avenue for enhancing FPAD and its generalization. Considering the fact that sensors are now being incorporated into mobile devices, there is still room for further exploration of multi-modal FPAD in order to enhance generalisation.

Advancing FPAD with the Power of Vision Transformers

The Vision Transformer (ViT) model [239] is a deep learning model that is based on the Transformer architecture. It was originally developed for Natural Language Processing (NLP) applications but has recently been widely applied to Computer Vision tasks. The introduction of ViT for image classification has also had a significant impact on the FPAD research community. A ViT model can extract meaningful features from a face image or video by treating the input as a series of tokens and processing them using the Transformer architecture. A significant amount of research has been conducted recently using ViT to address PA detection in FR systems [240, 183, 241]. The authors of [242] fine-tuned ViTs with FAS datasets, and compared their cross-dataset performance with other deep models, including ResNet-101 and DenseNet-161. According to the results, ViTs have the potential to improve FPAD performance and generalisation. It is important to note, however, that vision transformers require a greater amount of computational power as compared to deep pre-trained models. It is for this reason that the authors of [243] used the lightweight but efficient transformer model MobileViT [244] in order to address FPAD. The Mobile ViT module combines convolution and transformer functions to capture local and global information in an efficient manner. As a result, this shallow model can be used effectively as a visual translator on edge devices. In this regard, ViT models represent a high potential for future FPAD research.

Redefining Security: Generative AI and LLMs in FPAD

The influence of Large Language Models (LLM) and generative AI has extended beyond natural language processing tasks to contribute to vision-based tasks [245, 246]. As a result, advances in computer vision and multimodal AI have been made due to their ability to enable cross-modal understanding, data augmentation, content manipulation, and enhanced search capabilities, among other uses. Vision-based tasks continue to be developed and innovated as a result of these technologies. There is potential for them to be used both by attackers and by defenders in the fields of security and biometric identification [247, 248]. Recently, natural language-inspired processing was used to improve FPAD [249]. It was also shown that vision language pre-trained (VLP) models could improve feature learning to achieve generalisation [149] suggesting the possibility of further improvement in FPAD by using prompt engineering. As a result of the recent advancements, it is important to investigate how Generative AI and Large Language Models are altering the world of security [250], particularly in FPAD.

Crafting More Diversified Data Sets with Synthetic Samples

The variance present in a dataset is of utmost importance when it comes to the performance and generalization of FPAD. In an ideal scenario, the training dataset should display variance in various aspects such as attacks, background settings, illumination, size, ethnicity, gender, and resolution. However, the existing public datasets are limited in their variance of these factors, which in turn affects the FPAD performance negatively. Recent research [251] has demonstrated that a more diverse dataset can be created using image synthesis, which can help overcome the imbalances present in gender [252, 253], ethnicity [254, 255], and fairness [256]. Synthetic visual data creation techniques [257] can aid in creating artificial modalities and address issues such as privacy concerns.

Fusion: Exploring New Frontiers

The thesis proposed a methodology that combined deep features with hand-crafted features, but there are additional possibilities for exploring fusion in the context of FPAD. For instance, task-specific features can be extracted from RGB datasets using pre-trained models, while domain-specific features can be extracted from corresponding images of other modalities. The pre-trained models would need to be fine-tuned appropriately. By combining these different cues, fusion has the potential to improve both PA detection and generalization. Additionally, fusion can be extended beyond pre-trained deep models by combining them with vision transformers.

7.3 Conclusion

As new and advanced presentation attacks continue to emerge, it is crucial to create face presentation attack detection techniques that are both efficient and adaptable, in order to maintain the credibility of face recognition systems in security applications. This thesis demonstrates that by leveraging deep learning models to extract appropriate features, it is possible to substantially improve the generalization of attack detection. Moreover, maximizing the variability of the training dataset is a critical factor in determining the performance of the detection method. In this way, the power of deep learning can be harnessed to unmask imposters in front of security systems.

Bibliography

[1] Abdullakutty F, Elyan E, Johnston P. A review of state-of-the-art in Face Presentation Attack Detection: From early development to advanced deep learning and multi-modal fusion methods. Information fusion. 2021;75:55-69.

[2] Ross A, Banerjee S, Chowdhury A. Security in smart cities: A brief review of digital forensic schemes for biometric data. Pattern Recognition Letters. 2020;138:346-54.

[3] Anjos A, Komulainen J, Marcel S, Hadid A, Pietikäinen M. Face anti-spoofing: Visual approach. In: Handbook of biometric anti-spoofing. Springer; 2014. p. 65-82.

[4] Komulainen J, Boulkenafet Z, Akhtar Z. Review of face presentation attack detection competitions. In: Handbook of Biometric Anti-Spoofing. Springer; 2019. p. 291-317.

[5] Ramachandra R, Busch C. Presentation attack detection methods for face recognition systems: A comprehensive survey. ACM Computing Surveys (CSUR). 2017;50(1):1-37.

[6] Galbally J, Marcel S, Fierrez J. Biometric antispoofing methods: A survey in face recognition. IEEE Access. 2014;2:1530-52.

[7] Jia S, Guo G, Xu Z. A survey on 3D mask presentation attack detection and countermeasures. Pattern Recognition. 2020;98:107032.

[8] Boulkenafet Z, Komulainen J, Hadid A. Face anti-spoofing based on color texture analysis. In: 2015 IEEE international conference on image processing (ICIP). IEEE; 2015. p. 2636-40.

[9] Hasan MR, Mahmud SH, Li XY. Face Anti-Spoofing Using Texture-Based Techniques and Filtering Methods. In: Journal of Physics: Conference Series. vol. 1229. IOP Publishing; 2019. p. 012044.

[10] Tan X, Li Y, Liu J, Jiang L. Face liveness detection from a single image with sparse low rank bilinear discriminative model. In: European Conference on Computer Vision. Springer; 2010. p. 504-17.

[11] Abdullakutty F, Elyan E, Johnston P, Ali-Gombe A. Deep Transfer Learning on the Aggregated Dataset for Face Presentation Attack Detection. Cognitive computation. 2022;14(6):2223-33.

[12] Abdullakutty F, Johnston P, Elyan E. Fusion Methods for Face Presentation Attack Detection. Sensors. 2022;22(14):5196.

[13] Jia S, Guo G, Xu Z, Wang Q. Face presentation attack detection in mobile scenarios: A comprehensive evaluation. Image and Vision Computing. 2020;93:103826.

[14] Costa-Pazo A, Jiménez-Cabello D, Vázquez-Fernández E, Alba-Castro JL, López-Sastre RJ. Generalized presentation attack detection: a face anti-spoofing evaluation proposal. In: 2019 International Conference on Biometrics (ICB). IEEE; 2019. p. 1-8.

[15] Sharifi O. Face Anti-Spoofing Scheme Using Handcraft Based and Deep Learning Methods. Çukurova Üniversitesi Mühendislik-Mimarlık Fakültesi Dergisi;35(4):1103-10.

[16] Saad W, Shalaby WA, Shokair M, El-Samie FA, Dessouky M, Abdellatef E. COVID-19 classification using deep feature concatenation technique. Journal of Ambient Intelligence and Humanized Computing. 2022;13(4):2025-43.

[17] Kotwal K, Bhattacharjee S, Abbet P, Mostaani Z, Wei H, Wenkang X, et al. Domain-Specific Adaptation of CNN for Detecting Face Presentation Attacks in NIR. IEEE Transactions on Biometrics, Behavior, and Identity Science. 2022.

[18] LeCun Y, Bengio Y, Hinton G. Deep learning. nature. 2015;521(7553):436-44.

[19] Sinha RK, Pandey R, Pattnaik R. Deep learning for computer vision tasks: a review. arXiv preprint arXiv:180403928. 2018.

[20] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 770-8.

[21] Zhang X, Zou J, He K, Sun J. Accelerating very deep convolutional networks for classification and detection. IEEE transactions on pattern analysis and machine intelligence. 2015;38(10):1943-55.

[22] Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 2818-26.

[23] Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. Ieee; 2009. p. 248-55.

[24] Dhamecha TI, Nigam A, Singh R, Vatsa M. Disguise detection and face recognition in visible and thermal spectrums. In: 2013 International Conference on Biometrics (ICB). IEEE; 2013. p. 1-8.

[25] Kohli N, Yadav D, Noore A. Face verification with disguise variations via deep disguise recognizer. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops; 2018. p. 17-24.

[26] Kotwal K, Mostaani Z, Marcel S. Detection of age-induced makeup attacks on face recognition systems using multi-layer deep features. IEEE Transactions on Biometrics, Behavior, and Identity Science. 2019;2(1):15-25.

[27] Singh R, Agarwal A, Singh M, Nagpal S, Vatsa M. On the robustness of face recognition algorithms against attacks and bias. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34; 2020. p. 13583-9.

[28] Kisku DR, Rakshit RD. Face Spoofing and Counter-Spoofing: A Survey of State-of-the-art Algorithms. Transactions on Machine Learning and Artificial Intelligence. 2017;5(2):31-1.

[29] Zhang B, Tondi B, Barni M. Adversarial examples for replay attacks against CNN-based face recognition with anti-spoofing capability. Computer Vision and Image Understanding. 2020;197:102988.

[30] Deb D, Liu X, Jain AK. FaceGuard: A Self-Supervised Defense Against Adversarial Face Images. arXiv e-prints. 2020:arXiv-2011.

[31] Yang T, Zhao X, Wang X, Lv H. Evaluating facial recognition web services with adversarial and synthetic samples. Neurocomputing. 2020;406:378-85.

[32] Sharif M, Bhagavatula S, Bauer L, Reiter MK. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In: Proceedings of the 2016 acm sigsac conference on computer and communications security; 2016. p. 1528-40.

[33] Sharif M, Bhagavatula S, Bauer L, Reiter MK. A general framework for adversarial examples with objectives. ACM Transactions on Privacy and Security (TOPS). 2019;22(3):1-30.

[34] Zhou Z, Tang D, Wang X, Han W, Liu X, Zhang K. Invisible Mask: Practical Attacks on Face Recognition with Infrared. arXiv e-prints. 2018:arXiv-1803.

[35] Nguyen DL, Arora SS, Wu Y, Yang H. Adversarial Light Projection Attacks on Face Recognition Systems: A Feasibility Study. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops; 2020. p. 814-5.

[36] Singh M, Singh R, Ross A. A comprehensive overview of biometric fusion. Information Fusion. 2019;52:187-205.

[37] Tolosana R, Gomez-Barrero M, Busch C, Ortega-Garcia J. Biometric presentation attack detection: Beyond the visible spectrum. IEEE Transactions on Information Forensics and Security. 2019;15:1261-75.

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

[39] Souza L, Oliveira L, Pamplona M, Papa J. How far did we get in face spoofing detection? Engineering Applications of Artificial Intelligence. 2018;72:368-81.

[40] Ramachandra R, Stokkenes M, Mohammadi A, Venkatesh S, Raja K, Wasnik P, et al. Smartphone Multi-modal Biometric Authentication: Database and Evaluation. arXiv preprint arXiv:191202487. 2019.

[41] Zhang Z, Yan J, Liu S, Lei Z, Yi D, Li SZ. A face antispoofing database with diverse attacks. In: 2012 5th IAPR international conference on Biometrics (ICB). IEEE; 2012. p. 26-31.

[42] Hernandez-Ortega J, Fierrez J, Morales A, Galbally J. Introduction to face presentation attack detection. In: Handbook of Biometric Anti-Spoofing. Springer; 2019. p. 187-206.

[43] Wen D, Han H, Jain AK. Face spoof detection with image distortion analysis. IEEE Transactions on Information Forensics and Security. 2015;10(4):746-61.

[44] Ramachandra R, Venkatesh S, Raja KB, Bhattacharjee S, Wasnik P, Marcel S, et al. Custom silicone Face Masks: Vulnerability of Commercial Face Recognition Systems & Presentation Attack Detection. In: 2019 7th International Workshop on Biometrics and Forensics (IWBF). IEEE; 2019. p. 1-6.

[45] Bhattacharjee S, Mohammadi A, Marcel S. Spoofing deep face recognition with custom silicone masks. In: 2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS). IEEE; 2018. p. 1-7.

[46] Chingovska I, Anjos A, Marcel S. On the effectiveness of local binary patterns in face anti-spoofing. In: 2012 BIOSIG-proceedings of the international conference of biometrics special interest group (BIOSIG). IEEE; 2012. p. 1-7.

[47] Määttä J, Hadid A, Pietikäinen M. Face spoofing detection from single images using micro-texture analysis. In: 2011 international joint conference on Biometrics (IJCB). IEEE; 2011. p. 1-7.

[48] Komulainen J, Hadid A, Pietikäinen M. Context based face anti-spoofing. In: 2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS). IEEE; 2013. p. 1-8.

[49] Yang J, Lei Z, Liao S, Li SZ. Face liveness detection with component dependent descriptor. In: 2013 International Conference on Biometrics (ICB). IEEE; 2013. p. 1-6.

[50] Boulkenafet Z, Komulainen J, Hadid A. Face antispoofing using speeded-up robust features and fisher vector encoding. IEEE Signal Processing Letters. 2016;24(2):141-5.

[51] Peixoto B, Michelassi C, Rocha A. Face liveness detection under bad illumination conditions. In: 2011 18th IEEE International Conference on Image Processing. IEEE; 2011. p. 3557-60.

[52] Kumar S, Singh S, Kumar J. A comparative study on face spoofing attacks. In: 2017 International Conference on Computing, Communication and Automation (ICCCA). IEEE; 2017. p. 1104-8.

[53] Kaur R, Mann P. Techniques of face spoof detection: a review. Int J Comput Appl. 2017;164(1):29-33.

[54] Yılmaz AG, Turhal U, Nabiyev VV. EFFECT OF FEATURE SELECTION WITH META-HEURISTIC OPTIMIZATION METHODS ON FACE SPOOFING DETECTION. Journal of Modern Technology and Engineering. 2020;5(1):48-59.

[55] Patel K, Han H, Jain AK. Secure face unlock: Spoof detection on smartphones. IEEE transactions on information forensics and security. 2016;11(10):2268-83.

[56] Jourabloo A, Liu Y, Liu X. Face de-spoofing: Anti-spoofing via noise modeling. In: Proceedings of the European Conference on Computer Vision (ECCV); 2018. p. 290-306.

[57] Alotaibi A, Mahmood A. Deep face liveness detection based on nonlinear diffusion using convolution neural network. Signal, Image and Video Processing. 2017;11(4):713-20.

[58] Tirunagari S, Poh N, Windridge D, Iorliam A, Suki N, Ho AT. Detection of face spoofing using visual dynamics. IEEE transactions on information forensics and security. 2015;10(4):762-77.

[59] Arora G, Tiwari K, Gupta P. Liveness and Threat Aware Selfie Face Recognition. In: Selfie Biometrics. Springer; 2019. p. 197-210.

[60] Singh M, Arora A. A novel face liveness detection algorithm with multiple liveness indicators. Wireless Personal Communications. 2018;100(4):1677-87.

[61] Killioğlu M, Taşkiran M, Kahraman N. Anti-spoofing in face recognition with liveness detection using pupil tracking. In: 2017 IEEE 15th International Symposium on Applied Machine Intelligence and Informatics (SAMI). IEEE; 2017. p. 000087-92.

[62] Hernandez-Ortega J, Fierrez J, Morales A, Tome P. Time analysis of pulse-based face anti-spoofing in visible and NIR. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops; 2018. p. 544-52.

[63] Li X, Komulainen J, Zhao G, Yuen PC, Pietikäinen M. Generalized face anti-spoofing by detecting pulse from face videos. In: 2016 23rd International Conference on Pattern Recognition (ICPR). IEEE; 2016. p. 4244-9.

[64] Wang SY, Yang SH, Chen YP, Huang JW. Face liveness detection based on skin blood flow analysis. symmetry. 2017;9(12):305.

[65] Lin B, Li X, Yu Z, Zhao G. Face Liveness Detection by rPPG Features and Contextual Patch-Based CNN. In: Proceedings of the 2019 3rd International Conference on Biometric Engineering and Applications; 2019. p. 61-8.

[66] Gogate M, Dashtipour K, Adeel A, Hussain A. CochleaNet: A robust language-independent audio-visual model for real-time speech enhancement. Information Fusion. 2020;63:273 285.

[67] Adeel A, Gogate M, Hussain A, Whitmer WM. Lip-Reading Driven Deep Learning Approach for Speech Enhancement. IEEE Transactions on Emerging Topics in Computational Intelligence. 2019:1-10.

[68] Mahmud M, Kaiser MS, Hussain A, Vassanelli S. Applications of Deep Learning and Reinforcement Learning to Biological Data. IEEE Transactions on Neural Networks and Learning Systems. 2018;29(6):2063-79.

[69] Ieracitano C, Adeel A, Morabito FC, Hussain A. A novel statistical analysis and autoencoder driven intelligent intrusion detection approach. Neurocomputing. 2020;387:51 62.

[70] Pouyanfar S, Sadiq S, Yan Y, Tian H, Tao Y, Reyes MP, et al. A survey on deep learning: Algorithms, techniques, and applications. ACM Computing Surveys (CSUR). 2018;51(5):1-36.

[71] Rehman YAU, Po LM, Liu M. LiveNet: Improving features generalization for face liveness detection using convolution neural networks. Expert Systems with Applications. 2018;108:159 169.

[72] Chen FM, Wen C, Xie K, Wen FQ, Sheng GQ, Tang XG. Face liveness detection: fusing colour texture feature and deep feature. IET Biometrics. 2019;8(6):369-77.

[73] Yosinski J, Clune J, Bengio Y, Lipson H. How transferable are features in deep neural networks? In: Advances in neural information processing systems; 2014. p. 3320-8.

[74] Nagpal C, Dubey SR. A performance evaluation of convolutional neural networks for face anti spoofing. In: 2019 International Joint Conference on Neural Networks (IJCNN). IEEE; 2019. p. 1-8.

[75] Lucena O, Junior A, Moia V, Souza R, Valle E, Lotufo R. Transfer learning using convolutional neural networks for face anti-spoofing. In: International Conference Image Analysis and Recognition. Springer; 2017. p. 27-34.

[76] Yang J, Lei Z, Yi D, Li SZ. Person-specific face antispoofing with subject domain adaptation. IEEE Transactions on Information Forensics and Security. 2015;10(4):797-809.

[77] Nikisins O, George A, Marcel S. Domain Adaptation in Multi-Channel Autoencoder based Features for Robust Face Anti-Spoofing. In: International Conference on Biometrics 2019, IEEE. CONF; 2019. .

[78] Yu Z, Wan J, Qin Y, Li X, Li S, Zhao G. NAS-FAS: Static-Dynamic Central Difference Network Search for Face Anti-Spoofing. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2020.

[79] Wang G, Han H, Shan S, Chen X. Improving Cross-database Face Presentation Attack Detection via Adversarial Domain Adaptation. In: International Conference on Biometrics (ICB); 2019. .

[80] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative Adversarial Nets. In: Ghahramani Z, Welling M, Cortes C, Lawrence ND, Weinberger KQ, editors. Advances in Neural Information Processing Systems 27. Curran Associates, Inc.; 2014. p. 2672-80.

[81] Wang G, Han H, Shan S, Chen X. Unsupervised adversarial domain adaptation for cross-domain face presentation attack detection. IEEE Transactions on Information Forensics and Security. 2020;16:56-69.

[82] Zhou F, Gao C, Chen F, Li C, Li X, Yang F, et al. Face Anti-Spoofing Based on Multi-layer Domain Adaptation. In: 2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). IEEE; 2019. p. 192-7.

[83] Mohammadi A, Bhattacharjee S, Marcel S. Domain Adaptation for Generalization of Face Presentation Attack Detection in Mobile Settengs with Minimal Information. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE; 2020. p. 1001-5.

[84] Saha S, Xu W, Kanakis M, Georgoulis S, Chen Y, Paudel DP, et al. Domain Agnostic Feature Learning for Image and Video Based Face Anti-spoofing. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE; 2020. p. 3490-9.

[85] Wang G, Han H, Shan S, Chen X. Cross-domain face presentation attack detection via multi-domain disentangled representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020. p. 6678-87.

[86] Shao R, Lan X, Li J, Yuen PC. Multi-adversarial discriminative deep domain generalization for face presentation attack detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2019. p. 10023-31.

[87] Jia Y, Zhang J, Shan S, Chen X. Single-Side Domain Generalization for Face Anti-Spoofing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020. p. 8484-93.

[88] Zhang KY, Yao T, Zhang J, Tai Y, Ding S, Li J, et al. Face anti-spoofing via disentangled representation learning. In: European Conference on Computer Vision. Springer; 2020. p. 641-57.

[89] Arashloo SR, Kittler J, Christmas W. An anomaly detection approach to face spoofing detection: A new formulation and evaluation protocol. IEEE Access. 2017;5:13868-82.

[90] Arashloo SR, Kittler J. Client-Specific Anomaly Detection for Face Presentation Attack Detection. arXiv preprint arXiv:180700848. 2018.

[91] Nikisins O, Mohammadi A, Anjos A, Marcel S. On effectiveness of anomaly detection approaches against unseen presentation attacks in face anti-spoofing. In: 2018 International Conference on Biometrics (ICB). IEEE; 2018. p. 75-81.

[92] Fatemifar S, Arashloo SR, Awais M, Kittler J. Spoofing attack detection by anomaly detection. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE; 2019. p. 8464-8.

[93] Fatemifar S, Awais M, Arashloo SR, Kittler J. Combining multiple one-class classifiers for anomaly based face spoofing attack detection. In: 2019 International Conference on Biometrics (ICB). IEEE; 2019. p. 1-7.

[94] Pérez-Cabo D, Jiménez-Cabello D, Costa-Pazo A, López-Sastre RJ. Deep anomaly detection for generalized face anti-spoofing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops; 2019. p. 0-0.

[95] Abduh L, Ivrissimtzis I. Use of in-the-wild images for anomaly detection in face anti-spoofing. arXiv. 2020:arXiv-2006.

[96] Li Z, Li H, Lam KY, Kot AC. Unseen Face Presentation Attack Detection with Hypersphere Loss. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE; 2020. p. 2852-6.

[97] Feng H, Hong Z, Yue H, Chen Y, Wang K, Han J, et al. Learning Generalized Spoof Cues for Face Anti-spoofing. arXiv preprint arXiv:200503922. 2020.

[98] Baweja Y, Oza P, Perera P, Patel VM. Anomaly detection-based unknown face presentation attack detection. In: 2020 IEEE International Joint Conference on Biometrics (IJCB). IEEE; 2020. p. 1-9.

[99] Wang Y, Yao Q, Kwok JT, Ni LM. Generalizing from a few examples: A survey on few-shot learning. ACM Computing Surveys (CSUR). 2020;53(3):1-34.

[100] Qin Y, Zhao C, Zhu X, Wang Z, Yu Z, Fu T, et al. Learning meta model for zero-and few-shot face antispoofing. Association for Advancement of Artificial Intelligence (AAAI). 2020.

[101] Liu Y, Stehouwer J, Jourabloo A, Liu X. Deep tree learning for zero-shot face anti-spoofing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2019. p. 4680-9.

[102] Liu* Y, Jourabloo* A, Liu X. Learning Deep Models for Face Anti-Spoofing: Binary or Auxiliary Supervision. In: In Proceeding of IEEE Computer Vision and Pattern Recognition. Salt Lake City, UT; 2018. .

[103] Atoum Y, Liu Y, Jourabloo A, Liu X. Face anti-spoofing using patch and depth-based CNNs. In: 2017 IEEE International Joint Conference on Biometrics (IJCB). IEEE; 2017. p. 319-28.

[104] George A, Marcel S. Deep pixel-wise binary supervision for face presentation attack detection. In: 2019 International Conference on Biometrics (ICB). IEEE; 2019. p. 1-8.

[105] Wang Z, Zhao C, Qin Y, Zhou Q, Qi G, Wan J, et al. Exploiting temporal and depth information for multi-frame face anti-spoofing. arXiv preprint arXiv:181105118. 2018.

[106] Yu Z, Zhao C, Wang Z, Qin Y, Su Z, Li X, et al. Searching central difference convolutional networks for face anti-spoofing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020. p. 5295-305.

[107] Yu Z, Qin Y, Li X, Wang Z, Zhao C, Lei Z, et al. Multi-Modal Face Anti-Spoofing Based on Central Difference Networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops; 2020. p. 650-1.

[108] Yu Z, Li X, Niu X, Shi J, Zhao G. Face anti-spoofing with human material perception. In: European Conference on Computer Vision. Springer; 2020. p. 557-75.

[109] Kim T, Kim Y, Kim I, Kim D. BASN: Enriching Feature Representation Using Bipartite Auxiliary Supervisions for Face Anti-Spoofing. In: Proceedings of the IEEE International Conference on Computer Vision Workshops; 2019. p. 0-0.

[110] Liu Y, Stehouwer J, Jourabloo A, Atoum Y, Liu X. Presentation Attack Detection for Face in Mobile Phones. In: Selfie Biometrics. Springer; 2019. p. 171-96.

[111] Rehman YAU, Po LM, Liu M. SLNet: Stereo face liveness detection via dynamic disparity-maps and convolutional neural network. Expert Systems with Applications. 2020;142:113002.

[112] Yang X, Luo W, Bao L, Gao Y, Gong D, Zheng S, et al. Face anti-spoofing: Model matters, so does data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2019. p. 3507-16.

[113] Munir R, Khan RA. An extensive review on spectral imaging in biometric systems: Challenges & advancements. Journal of Visual Communication and Image Representation. 2019;65:102660.

[114] Jiang F, Liu P, Zhou X. Multilevel fusing paired visible light and near-infrared spectral images for face anti-spoofing. Pattern Recognition Letters. 2019;128:30 37.

[115] George A, Mostaani Z, Geissenbuhler D, Nikisins O, Anjos A, Marcel S. Biometric Face Presentation Attack Detection With Multi-Channel Convolutional Neural Network. IEEE Transactions on Information Forensics and Security. 2020;15:42-55.

[116] Kotwal K, Bhattacharjee S, Marcel S. Multispectral Deep Embeddings as a Countermeasure to Custom Silicone Mask Presentation Attacks. IEEE Transactions on Biometrics, Behavior, and Identity Science. 2019 Oct;1(4):238-51.

[117] Chen H, Hu G, Lei Z, Chen Y, Robertson NM, Li SZ. Attention-Based Two-Stream Convolutional Networks for Face Spoofing Detection. IEEE Transactions on Information Forensics and Security. 2019;15:578-93.

[118] Fan Y, Shi Y, Wang X, Yi H. Research on Liveness Detection Algorithms Based on Deep Learning. In: 2019 IEEE 10th International Conference on Software Engineering and Service Science (ICSESS). IEEE; 2019. p. 1-6.

[119] Wang G, Lan C, Han H, Shan S, Chen X. Multi-Modal Face Presentation Attack Detection via Spatial and Channel Attentions. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). IEEE Computer Society; 2019. p. 1584-90.

[120] Jiang F, Liu P, Shao X, Zhou X. Face anti-spoofing with generated near-infrared images. Multimedia Tools and Applications. 2020:1-25.

[121] Liu S, Song Y, Zhang M, Zhao J, Yang S, Hou K. An Identity Authentication Method Combining Liveness Detection and Face Recognition. Sensors. 2019;19(21):4733.

[122] Hong Y, Hwang U, Yoo J, Yoon S. How generative adversarial networks and their variants work: An overview. ACM Computing Surveys (CSUR). 2019;52(1):1-43.

[123] Li L, Feng X, Xia Z, Jiang X, Hadid A. Face spoofing detection with local binary pattern network. Journal of visual communication and image representation. 2018;54:182-92.

[124] Thippeswamy G, Vinutha H, Dhanapal R. A New Ensemble of Texture Descriptors Based On Local Appearance-Based Methods For Face Anti-Spoofing System. J Crit Rev. 2020;7(11):644-9.

[125] Liu W, Wei X, Lei T, Wang X, Meng H, Nandi AK. Data Fusion based Two-stage Cascade Framework for Multi-Modality Face Anti-Spoofing. IEEE Transactions on Cognitive and Developmental Systems. 2021.

[126] Huang X, Huang Q, Zhang N. Dual fusion paired environmental background and face region for face anti-spoofing. In: 2021 5th Asian Conference on Artificial Intelligence Technology (ACAIT). IEEE; 2021. p. 142-9.

[127] Younis MC, Abuhammad H. A hybrid fusion framework to multi-modal bio metric identification. Multimedia Tools and Applications. 2021:1-24.

[128] Fang M, Damer N, Kirchbuchner F, Kuijper A. Learnable Multi-level Frequency Decomposition and Hierarchical Attention Mechanism for Generalized Face Presentation Attack Detection. arXiv preprint arXiv:210907950. 2021.

[129] Edwards T, Hossain MS. Effectiveness of Deep Learning on Serial Fusion Based Biometric Systems. IEEE Transactions on Artificial Intelligence. 2021.

[130] Daniel N, Anitha A. Texture and quality analysis for face spoofing detection. Computers & Electrical Engineering. 2021;94:107293.

[131] Xu Y, Wang Z, Han H, Wu L, Liu Y. Exploiting Non-uniform Inherent Cues to Improve Presentation Attack Detection. In: 2021 IEEE International Joint Conference on Biometrics (IJCB). IEEE; 2021. p. 1-8.

[132] Chen B, Yang W, Li H, Wang S, Kwong S. Camera Invariant Feature Learning for Generalized Face Anti-spoofing. IEEE Transactions on Information Forensics and Security. 2021;16:2477-92.

[133] Cai R, Li Z, Wan R, Li H, Hu Y, Kot AC. Learning Meta Pattern for Face Anti-Spoofing. arXiv preprint arXiv:211006753. 2021.

[134] Song X, Wu Q, Yu D, Hu G, Wu X. Face Anti-Spoofing Detection Using Least Square Weight Fusion of Channel-Based Feature Classifiers. EasyChair; 2020.

[135] Anand A, Vishwakarma DK. Face Anti-Spoofing by Spatial Fusion of Colour Texture Features and Deep Features. In: 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS). IEEE; 2020. p. 1012-7.

[136] Chen S, Li W, Yang H, Huang D, Wang Y. 3D Face Mask Anti-spoofing via Deep Fusion of Dynamic Texture and Shape Clues. In: 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020). IEEE; 2020. p. 314-21.

[137] Muhammad U, Hoque MZ, Oussalah M, Laaksonen J. Deep ensemble learning with frame skipping for face anti-spoofing. arXiv preprint arXiv:230702858. 2023.

[138] Amuthavalli S, Uma Kumari C. Computational Analysis and Performance Investigation of Convolutional Neural Network-Based Algorithms for Effective Face Spoof Detection. In: International Conference on Data Analytics and Insights. Springer; 2023. p. 481-90.

[139] Boutros F, Struc V, Fierrez J, Damer N. Synthetic data for face recognition: Current state and future prospects. Image and Vision Computing. 2023:104688.

[140] Muhammad U, Beddiar DR, Oussalah M. Domain Generalization via Ensemble Stacking for Face Presentation Attack Detection. arXiv preprint arXiv:230102145. 2023.

[141] Ibsen M, Rathgeb C, Brechtel F, Klepp R, Pöppelmann K, George A, et al. Attacking Face Recognition with T-shirts: Database, Vulnerability Assessment and Detection. IEEE Access. 2023.

[142] Alkhunaizi N, Srivatsan K, Almalik F, Almakky I, Nandakumar K. FedSIS: Federated Split Learning with Intermediate Representation Sampling for Privacy-preserving Generalized Face Presentation Attack Detection. arXiv preprint arXiv:230810236. 2023.

[143] Solomon E, Cios KJ. FASS: Face Anti-Spoofing System Using Image Quality Features and Deep Learning. Electronics. 2023;12(10):2199.

[144] Shu X, Li X, Zuo X, Xu D, Shi J. Face spoofing detection based on multi-scale color inversion dual-stream convolutional neural network. Expert Systems with Applications. 2023;224:119988.

[145] Chang CJ, Lee YC, Yao SH, Chen MH, Wang CY, Lai SH, et al. A Closer Look at Geometric Temporal Dynamics for Face Anti-Spoofing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2023. p. 1081-91.

[146] Kong C, Zheng K, Liu Y, Wang S, Rocha A, Li H. M3FAS: An Accurate and Robust MultiModal Mobile Face Anti-Spoofing System. arXiv preprint arXiv:230112831. 2023.

[147] Yu Z, Liu A, Zhao C, Cheng KH, Cheng X, Zhao G. Flexible-modal face anti-spoofing: A benchmark. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2023. p. 6345-50.

[148] Muhammad U, Oussalah M, Hoque MZ, Laaksonen J. Saliency-based video summarization for face anti-spoofing. arXiv preprint arXiv:230812364. 2023.

[149] Srivatsan K, Naseer M, Nandakumar K. FLIP: Cross-domain Face Anti-spoofing with Language Guidance. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2023. p. 19685-96.

[150] Wang Z, Xu Y, Wu L, Han H, Ma Y, Li Z. Improving Face Anti-spoofing via Advanced Multi-perspective Feature Learning. ACM Transactions on Multimedia Computing, Communications and Applications. 2023;19(6):1-18.

[151] Anjos A, Chingovska I, Marcel S. Anti-spoofing: Face databases. Springer US; 2014.

[152] Dantcheva A, Chen C, Ross A. Can facial cosmetics affect the matching accuracy of face recognition systems? In: 2012 IEEE Fifth international conference on biometrics: theory, applications and systems (BTAS). IEEE; 2012. p. 391-8.

[153] Bhattacharjee S, Marcel S. What you can't see can help you-extended-range imaging for 3d-mask presentation attack detection. In: 2017 International Conference of the Biometrics Special Interest Group (BIOSIG). IEEE; 2017. p. 1-7.

[154] Chen C, Dantcheva A, Ross A. Automatic facial makeup detection with application in face recognition. In: 2013 international conference on biometrics (ICB). IEEE; 2013. p. 1-8.

[155] Agarwal A, Yadav D, Kohli N, Singh R, Vatsa M, Noore A. Face presentation attack with latex masks in multispectral videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops; 2017. p. 81-9.

[156] Raghavendra R, Raja KB, Busch C. Presentation attack detection for face recognition using light field camera. IEEE Transactions on Image Processing. 2015;24(3):1060-75.

[157] Galbally J, Satta R. Three-dimensional and two-and-a-half-dimensional face recognition spoofing using three-dimensional printed models. IET Biometrics. 2016;5(2):83-91.

[158] Erdogmus N, Marcel S. Spoofing 2D face recognition systems with 3D masks. In: 2013 International Conference of the BIOSIG Special Interest Group (BIOSIG). IEEE; 2013. p. 1-8.

[159] Liu S, Yang B, Yuen PC, Zhao G. A 3D mask face anti-spoofing database with real world variations. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops; 2016. p. 100-6.

[160] Chingovska I, Erdogmus N, Anjos A, Marcel S. Face recognition systems under spoofing attacks. In: Face Recognition Across the Imaging Spectrum. Springer; 2016. p. 165-94.

[161] Costa-Pazo A, Bhattacharjee S, Vazquez-Fernandez E, Marcel S. The replay-mobile face presentation-attack database. In: 2016 International Conference of the Biometrics Special Interest Group (BIOSIG). IEEE; 2016. p. 1-7.

[162] Steiner H, Sporrer S, Kolb A, Jung N. Design of an active multispectral SWIR camera system for skin detection and face verification. Journal of Sensors. 2016;2016.

[163] Raghavendra R, Raja KB, Venkatesh S, Cheikh FA, Busch C. On the vulnerability of extended multispectral face recognition systems towards presentation attacks. In: 2017 IEEE International Conference on Identity, Security and Behavior Analysis (ISBA). IEEE; 2017. p. 1-8.

[164] Chen C, Dantcheva A, Swearingen T, Ross A. Spoofing faces using makeup: An investigative study. In: 2017 IEEE International Conference on Identity, Security and Behavior Analysis (ISBA). IEEE; 2017. p. 1-8.

[165] Boulkenafet Z, Komulainen J, Li L, Feng X, Hadid A. Oulu-npu: A mobile face presentation attack database with real-world variations. In: 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017). IEEE; 2017. p. 612-8.

[166] Manjani I, Tariyal S, Vatsa M, Singh R, Majumdar A. Detecting silicone mask-based presentation attack via deep dictionary learning. IEEE Transactions on Information Forensics and Security. 2017;12(7):1713-23.

[167] Singh M, Singh R, Vatsa M, Ratha NK, Chellappa R. Recognizing disguised faces in the wild. IEEE Transactions on Biometrics, Behavior, and Identity Science. 2019;1(2):97-108.

[168] Li H, Li W, Cao H, Wang S, Huang F, Kot AC. Unsupervised domain adaptation for face anti-spoofing. IEEE Transactions on Information Forensics and Security. 2018;13(7):1794-809.

[169] Xiao J, Tang Y, Guo J, Yang Y, Zhu X, Lei Z, et al. 3DMA: A Multi-modality 3D Mask Face Anti-spoofing Database. In: 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). IEEE; 2019. p. 1-8.

[170] Kotwal K, Mostaani Z, Marcel S. Detection of Age-Induced Makeup Attacks on Face Recognition Systems Using Multi-Layer Deep Features. IEEE Transactions on Biometrics, Behavior, and Identity Science. 2020 Jan;2(1):15-25.

[171] Zhang S, Liu A, Wan J, Liang Y, Guo G, Escalera S, et al. CASIA-SURF: A Large-scale Multi-modal Benchmark for Face Anti-spoofing. IEEE Transactions on Biometrics, Behavior, and Identity Science. 2020:1-1.

[172] Timoshenko D, Simonchik K, Shutov V, Zhelezneva P, Grishkin V. Large Crowd Collected Facial Anti-Spoofing Dataset. In: 2019 Computer Science and Information Technologies (CSIT). IEEE; 2019. p. 123-6.

[173] Bok JY, Suh KH, Lee EC. Verifying the Effectiveness of New Face Spoofing DB with Capture Angle and Distance. Electronics. 2020;9(4):661.

[174] Jia S, Li X, Hu C, Xu Z. Spoofing and Anti-Spoofing with Wax Figure Faces. arXiv preprint arXiv:191005457. 2019.

[175] Liu A, Tan Z, Wan J, Escalera S, Guo G, Li SZ. Casia-surf cefa: A benchmark for multi-modal cross-ethnicity face anti-spoofing. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision; 2021. p. 1179-87.

[176] Zhang Y, Yin Z, Li Y, Yin G, Yan J, Shao J, et al. CelebA-Spoof: Large-Scale Face Anti-Spoofing Dataset with Rich Annotations. In: European Conference on Computer Vision. Springer; 2020. p. 70-85.

[177] Kushwaha V, Singh M, Singh R, Vatsa M, Ratha N, Chellappa R. Disguised faces in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops; 2018. p. 1-9.

[178] Liu Y, Jourabloo A, Liu X. Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2018. p. 389-98.

[179] Dhamecha TI, Singh R, Vatsa M, Kumar A. Recognizing disguised faces: Human and machine evaluation. PloS one. 2014;9(7).

[180] Heusch G, George A, Geissbühler D, Mostaani Z, Marcel S. Deep Models and Shortwave Infrared Information to Detect Face Presentation Attacks. IEEE Transactions on Biometrics, Behavior, and Identity Science. 2020. Available from: http://publications.idiap.ch/downloads/papers/2020/Heusch_TBIOM_2020.pdf.

[181] Rostami M, Spinoulas L, Hussein M, Mathai J, Abd-Almageed W. Detection and continual learning of novel face presentation attacks. In: Proceedings of the IEEE/CVF international conference on computer vision; 2021. p. 14851-60.

[182] Kim Y, Gwak H, Oh J, Kang M, Kim J, Kwon H, et al. CloudNet: A LiDAR-Based Face Anti-Spoofing Model That Is Robust Against Light Variation. IEEE Access. 2023;11:16984-93.

[183] Zhang D, Meng J, Zhang J, Deng X, Ding S, Zhou M, et al. SonarGuard: Ultrasonic Face Liveness Detection on Mobile Devices. IEEE Transactions on Circuits and Systems for Video Technology. 2023.

[184] Fang H, Liu A, Wan J, Escalera S, Zhao C, Zhang X, et al. Surveillance Face Anti-spoofing. arXiv preprint arXiv:230100975. 2023.

[185] Nguyen HP, Delahaies A, Retraint F, Morain-Nicolier F. Face presentation attack detection based on a statistical model of image noise. IEEE Access. 2019;7:175429-42.

[186] Chingovska I, Dos Anjos AR, Marcel S. Biometrics evaluation under spoofing attacks. IEEE transactions on Information Forensics and Security. 2014;9(12):2264-76.

[187] Galbally J, Alonso-Fernandez F, Fierrez J, Ortega-Garcia J. A high performance fingerprint liveness detection method based on quality related features. Future Generation Computer Systems. 2012;28(1):311-21.

[188] Pan G, Sun L, Wu Z, Lao S. Eyeblink-based anti-spoofing in face recognition from a generic webcamera. In: 2007 IEEE 11th International Conference on Computer Vision. IEEE; 2007. p. 1-8.

[189] Information Technology- Biometric Presentation Attack Detection- Part 3: Testing and Reporting. International Organization for Standardization. 2017;ISO/IEC DIS 30107-3:2017.

[190] Bhattacharjee S, Mohammadi A, Anjos A, Marcel S. Recent advances in face presentation attack detection. In: Handbook of Biometric Anti-Spoofing. Springer; 2019. p. 207-28.

[191] Liu A, Li X, Wan J, Liang Y, Escalera S, Escalante HJ, et al. Cross-ethnicity face anti-spoofing recognition challenge: A review. IET Biometrics.

[192] Batzner K, Heckler L, König R. Efficientad: Accurate visual anomaly detection at millisecond-level latencies. arXiv preprint arXiv:230314535. 2023.

[193] Xiong F, AbdAlmageed W. Unknown presentation attack detection with face RGB images. In: 2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS). IEEE; 2018. p. 1-9.

[194] Rattani A, Derakhshani R. A survey of mobile face biometrics. Computers & Electrical Engineering. 2018;72:39-52.

[195] Peng F, Qin L, Long M. Face presentation attack detection based on chromatic co-occurrence of local binary pattern and ensemble learning. Journal of Visual Communication and Image Representation. 2020;66:102746.

[196] Abdullakutty F, Elyan E, Johnston P. Face Spoof Detection: An Experimental Framework. In: International Conference on Engineering Applications of Neural Networks. Springer; 2021. p. 293-304.

[197] Fahn CS, Lee CP, Wu ML. A Cross-Dataset Evaluation of Anti-Face-Spoofing Methods Using Random Forests and Convolutional Neural Networks. In: Proceedings of the 2019 2nd Artificial Intelligence and Cloud Computing Conference; 2019. p. 89-96.

[198] Angadi SA, Kagawade VC. Detection of face spoofing using multiple texture descriptors. In: 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS). IEEE; 2018. p. 151-6.

[199] Satapathy A, Livingston LJ. A lite convolutional neural network built on permuted Xceptio-inception and Xceptio-reduction modules for texture based facial liveness recognition. Multimedia Tools and Applications. 2020:1-32.

[200] Parveen S, Ahmad SMS, Hanafi M, Adnan WAW. Face anti-spoofing methods. Current science. 2015:1491-500.

[201] Luan X, Wang H, Ou W, Liu L. Face liveness detection with recaptured feature extraction. In: 2017 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC). IEEE; 2017. p. 429-32.

[202] Koshy R, Mahmood A. Optimizing deep CNN architectures for face liveness detection. Entropy. 2019;21(4):423.

[203] Parveen S, Ahmad SMS, Abbas NH, Adnan WAW, Hanafi M, Naeem N. Face liveness detection using dynamic local ternary pattern (DLTP). Computers. 2016;5(2):10.

[204] Beham MP, Roomi SMM. Anti-spoofing enabled face recognition based on aggregated local weighted gradient orientation. Signal, Image and Video Processing. 2018;12(3):531-8.

[205] Vanitha A, Vaidehi V, Vasuhi S. Liveliness detection in real time videos using color based chromatic moment feature. In: 2018 International Conference on Recent Trends in Advance Computing (ICRTAC). IEEE; 2018. p. 162-7.

[206] Şengür A, Akhtar Z, Akbulut Y, Ekici S, Budak Ü. Deep feature extraction for face liveness detection. In: 2018 International Conference on Artificial Intelligence and Data Processing (IDAP). IEEE; 2018. p. 1-4.

[207] Beham MP, Roomi SMM, Jebina H, Kavitha M. Face spoofing detection using binary gradient orientation pattern with deep neural network. In: 2017 Ninth International Conference on Advances in Pattern Recognition (ICAPR). IEEE; 2017. p. 1-6.

[208] Boulkenafet Z, Komulainen J, Hadid A. Face spoofing detection using colour texture analysis. IEEE Transactions on Information Forensics and Security. 2016;11(8):1818-30.

[209] Ojala T, Pietikainen M, Maenpaa T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on pattern analysis and machine intelligence. 2002;24(7):971-87.

[210] Stricker MA, Orengo M. Similarity of color images. In: Storage and retrieval for image and video databases III. vol. 2420. SPiE; 1995. p. 381-92.

[211] Costa-Pazo A, Pérez-Cabo D, Jiménez-Cabello D, Alba-Castro JL, Vazquez-Fernandez E. Face presentation attack detection. A comprehensive evaluation of the generalisation problem. IET Biometrics. 2021;10(4):408-29.

[212] Liu M, Mu J, Yu Z, Ruan K, Shu B, Yang J. Adversarial learning and decomposition-based domain generalization for face anti-spoofing. Pattern Recognition Letters. 2022;155:171-7.

[213] Shi L, Zhang J, Liang C, Shan S. Unknown Aware Feature Learning for Face Forgery Detection. In: 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021). IEEE; 2021. p. 1-5.

[214] Heusch G, George A, Geissbühler D, Mostaani Z, Marcel S. Deep models and shortwave infrared information to detect face presentation attacks. IEEE Transactions on Biometrics, Behavior, and Identity Science. 2020;2(4):399-409.

[215] Bresan R, Beluzo C, Carvalho T. Exposing Presentation Attacks by a Combination of Multi-intrinsic Image Properties, Convolutional Networks and Transfer Learning. In: International Conference on Advanced Concepts for Intelligent Vision Systems. Springer; 2020. p. 153-65.

[216] Tu X, Fang Y. Ultra-deep neural network for face anti-spoofing. In: International Conference on Neural Information Processing. Springer; 2017. p. 686-95.

[217] Huang GB, Mattar M, Berg T, Learned-Miller E. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In: Workshop on faces in'Real-Life'Images: detection, alignment, and recognition; 2008. .

[218] Kotwal K, Marcel S. CNN Patch Pooling for Detecting 3D Mask Presentation Attacks in NIR. In: 2020 IEEE International Conference on Image Processing (ICIP). IEEE; 2020. p. 1336-40.

[219] Li L, Gao Z, Huang L, Zhang H, Lin M. A Dual-Modal Face Anti-Spoofing Method via Light-Weight Networks. In: 2019 IEEE 13th International Conference on Anti-counterfeiting, Security, and Identification (ASID). IEEE; 2019. p. 70-4.

[220] Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 4700-8.

[221] Satapathy A, Livingston LJ. A lite convolutional neural network built on permuted Xceptio-inception and Xceptio-reduction modules for texture based facial liveness recognition. Multimedia Tools and Applications. 2021;80(7):10441-72.

[222] Koppikar U, Sujatha C, Patil P, Hiremath P. Face Liveness Detection to Overcome Spoofing Attacks in Face Recognition System. In: Innovations in Computational Intelligence and Computer Vision. Springer; 2021. p. 351-60.

[223] Hadiprakoso RB, Setiawan H, et al. Face Anti-Spoofing Using CNN Classifier & Face liveness Detection. In: 2020 3rd International Conference on Information and Communications Technology (ICOIACT). IEEE; 2020. p. 143-7.

[224] Wang Y, Song X, Xu T, Feng Z, Wu XJ. From RGB to Depth: Domain Transfer Network for Face Anti-Spoofing. IEEE Transactions on Information Forensics and Security. 2021.

[225] Sharma D, Selwal A. A face anti-spoofing approach based on generic sequential model using scale invariant features. In: 2021 13th International Conference on Electronics, Computers and Artificial Intelligence (ECAI). IEEE; 2021. p. 1-6.

[226] Chen B, Yang W, Wang S. Generalized Face Antispoofing by Learning to Fuse Features From High-and Low-Frequency Domains. IEEE MultiMedia. 2021;28(1):56-64.

[227] Zhang Z, Jiang C, Zhong X, Song C, Zhang Y. Two-stream Convolutional Networks for Multi-frame Face Anti-spoofing. arXiv preprint arXiv:210804032. 2021.

[228] Huang X, Xia J, Shen L. One-Class Face Anti-spoofing Based on Attention Auto-encoder. In: Chinese Conference on Biometric Recognition. Springer; 2021. p. 365-73.

[229] Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980. 2014.

[230] Hashemifard S, Akbari M. A Compact Deep Learning Model for Face Spoofing Detection. arXiv preprint arXiv:210104756. 2021.

[231] Edmunds T, Caplier A. Face spoofing detection based on colour distortions. IET biometrics. 2018;7(1):27-38.

[232] Abdullakutty F, Elyan E, Johnston P. Unmasking the Imposters: Task-specific feature learning for face presentation attack detection. In: 2023 International Joint Conference on Neural Networks (IJCNN). IEEE; 2023. p. 1-8.

[233] Kolkur S, Kalbande D. Survey of texture based feature extraction for skin disease detection. In: 2016 International Conference on ICT in Business Industry & Government (ICTBIG). IEEE; 2016. p. 1-6.

[234] Kim G, Eum S, Suhr JK, Kim DI, Park KR, Kim J. Face liveness detection based on texture and frequency analyses. In: 2012 5th IAPR international conference on biometrics (ICB). IEEE; 2012. p. 67-72.

[235] Wang C, Yu B, Zhou J. A Learnable Gradient operator for face presentation attack detection. Pattern Recognition. 2023;135:109146.

[236] Joseph VR, Vakayil A. SPlit: An optimal method for data splitting. Technometrics. 2022;64(2):166-76.

[237] Joseph VR. Optimal ratio for data splitting. Statistical Analysis and Data Mining: The ASA Data Science Journal. 2022;15(4):531-8.

[238] Lyu X, Ren C, Ni W, Tian H, Liu RP, Dutkiewicz E. Optimal online data partitioning for geo-distributed machine learning in edge of wireless networks. IEEE Journal on Selected Areas in Communications. 2019;37(10):2393-406.

[239] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:201011929. 2020.

[240] Yu Z, Cai R, Cui Y, Liu X, Hu Y, Kot A. Rethinking Vision Transformer and Masked Autoencoder in Multimodal Face Anti-Spoofing. arXiv preprint arXiv:230205744. 2023.

[241] Ming Z, Yu Z, Al-Ghadi M, Visani M, Luqman MM, Burie JC. Vitranspad: video transformer using convolution and self-attention for face presentation attack detection. In: 2022 IEEE International Conference on Image Processing (ICIP). IEEE; 2022. p. 4248-52.

[242] George A, Marcel S. On the effectiveness of vision transformers for zero-shot face anti-spoofing. In: 2021 IEEE International Joint Conference on Biometrics (IJCB). IEEE; 2021. p. 1-8.

[243] Liao CH, Chen WC, Liu HT, Yeh YR, Hu MC, Chen CS. Domain Invariant Vision Transformer Learning for Face Anti-Spoofing. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision; 2023. p. 6098-107.

[244] Mehta S, Rastegari M. Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer. arXiv preprint arXiv:211002178. 2021.

[245] Li Z, Wang Y, Du M, Liu Q, Wu B, Zhang J, et al. ReForm-Eval: Evaluating Large Vision Language Models via Unified Re-Formulation of Task-Oriented Benchmarks. arXiv preprint arXiv:231002569. 2023.

[246] Du H, Niyato D, Kang J, Xiong Z, Lam KY, Fang Y, et al. Spear or Shield: Leveraging Generative AI to Tackle Security Threats of Intelligent Network Services. arXiv preprint arXiv:230602384. 2023.

[247] Cardenuto JP, Yang J, Padilha R, Wan R, Moreira D, Li H, et al. The Age of Synthetic Realities: Challenges and Opportunities. arXiv preprint arXiv:230611503. 2023.

[248] Tiong LCO, Sigmund D, Teoh ABJ. Face-Periocular Cross-Identification via Contrastive Hybrid Attention Vision Transformer. IEEE Signal Processing Letters. 2023;30:254-8.

[249] Mirzaalian H, Hussein ME, Spinoulas L, May J, Abd-Almageed W. Explaining face presentation attack detection using natural language. In: 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021). IEEE; 2021. p. 1-8.

[250] Wang T, Zhang Y, Qi S, Zhao R, Xia Z, Weng J. Security and Privacy on Generative Data in AIGC: A Survey. arXiv preprint arXiv:230909435. 2023.

[251] Fang M, Huber M, Damer N. SynthASpoof: Developing Face Presentation Attack Detection Based on Privacy-friendly Synthetic Data. arXiv preprint arXiv:230302660. 2023.

[252] Alshareef N, Yuan X, Roy K, Atay M. A Study of Gender Bias in Face Presentation Attack and Its Mitigation. Future Internet. 2021;13(9):234.

[253] Sharma D, Selwal A. A survey on face presentation attack detection mechanisms: hitherto and future perspectives. Multimedia Systems. 2023:1-51.

[254] Abduh L, Ivrissimtzis I. Race Bias Analysis of Bona Fide Errors in Face Anti-spoofing. arXiv preprint arXiv:221005366. 2022.

[255] Yu Z, Komulainen J, Li X, Zhao G. Review of Face Presentation Attack Detection Competitions. In: Handbook of Biometric Anti-Spoofing: Presentation Attack Detection and Vulnerability Assessment. Springer; 2023. p. 287-336.

[256] Fang M, Yang W, Kuijper A, Struc V, Damer N. Fairness in Face Presentation Attack Detection. arXiv preprint arXiv:220909035. 2022.

[257] Makrushin A, Uhl A, Dittmann J. A Survey On Synthetic Biometrics: Fingerprint, Face, Iris And Vascular Patterns. IEEE Access. 2023.