Spring 2022

Learning Domain Invariant Information to Enhance Presentation Attack Detection in Visible Face Recognition Systems

Jennifer Hamblin
University of Nebraska-Lincoln, jennlhamblin@gmail.com

LEARNING DOMAIN INVARIANT INFORMATION TO ENHANCE PRESENTATION ATTACK DETECTION IN VISIBLE FACE RECOGNITION SYSTEMS

by

Jennifer Hamblin

A THESIS

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfilment of Requirements

For the Degree of Master of Science

Major: Electrical Engineering

Under the Supervision of Professor Benjamin Riggan

Lincoln, Nebraska

May, 2022

LEARNING DOMAIN INVARIANT INFORMATION TO ENHANCE PRESENTATION ATTACK DETECTION IN VISIBLE FACE RECOGNITION SYSTEMS

Jennifer Hamblin, M.S.

University of Nebraska, 2022

Adviser: Benjamin Riggan

Face signatures, including size, shape, texture, skin tone, eye color, appearance, and scars/marks, are widely used as discriminative, biometric information for access control. Despite recent advancements in facial recognition systems, presentation attacks on facial recognition systems have become increasingly sophisticated. The ability to detect presentation attacks or spoofing attempts is a pressing concern for the integrity, security, and trust of facial recognition systems. Multi-spectral imaging has been previously introduced as a way to improve presentation attack detection by utilizing sensors that are sensitive to different regions of the electromagnetic spectrum (e.g., visible, near infrared, long-wave infrared). Although multi-spectral presentation attack detection systems may be discriminative, the need for additional sensors and computational resources substantially increases complexity and costs. Instead, we propose a method that exploits information from infrared imagery during training to increase the discriminability of visible-based presentation attack detection systems. We introduce (1) a new cross-domain presentation attack detection framework that increases the separability of bonafide and presentation attacks using only visible spectrum imagery, (2) an inverse domain regularization technique for added training stability when optimizing our cross-domain presentation attack detection framework, and (3) a dense domain adaptation subnetwork to transform representations between

visible and non-visible domains.

ACKNOWLEDGMENTS

I want to give a huge thanks to my advisor Professor Benjamin Riggan, without his encouragement, patience, and guidance none of this work would have been possible.

The support of my thesis committee members, Professors Khalid Sayood and Michael Hoffman, was also crucial to my advancement through this program. I will always be grateful for their tolerance of my unannounced office visits.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Faces are among the most prevalent biometric modalities (face, iris, fingerprint, voice) which are used in consumer devices such as personal cell phone, tablet, and computer identity authentication, as well as in commercial security systems, for airport security, and at border crossings. One advantage of using faces for biometric recognition is that acquisition of facial imagery can be contactless, covert and non-intrusive. Currently, deep learning networks are nearing human level performance on face recognition tasks (Taigman et al. (2014); Guo and Zhang (2019)). However, the ubiquity of facial recognition systems and increasing vulnerabilities, such as identity spoofing or presentation attacks (PAs), necessitates enhanced security measures to prevent failures in enrollment, authentication, or identification.

PAs describe the process of altering or obscuring facial signatures to gain access or evade detection. We study the issue of presentation attack detection (PAD) using multiple imaging domains to enhance the information in visible imagery to better differentiate between genuine (bonafide) and attack samples.

1.1 Presentation Attack Detection

The primary objective of PAD is to equip biometric systems with the ability to identify intentional attacks on the system. Because PAs include instances where a person may be trying to avoid detection, it is necessary to build a system that is sensitive to intentional obfuscations of appearance. Face PAs can include complete obfuscations, such as printed image, video replay, or mask style attacks, or partial obfuscations, such as wearing glasses, make-up, or wigs. The partial obfuscations are generally more challenging to detect because these attacks are often acceptable societal behaviors and practices. Figure 1.1 shows several examples of PA instruments.

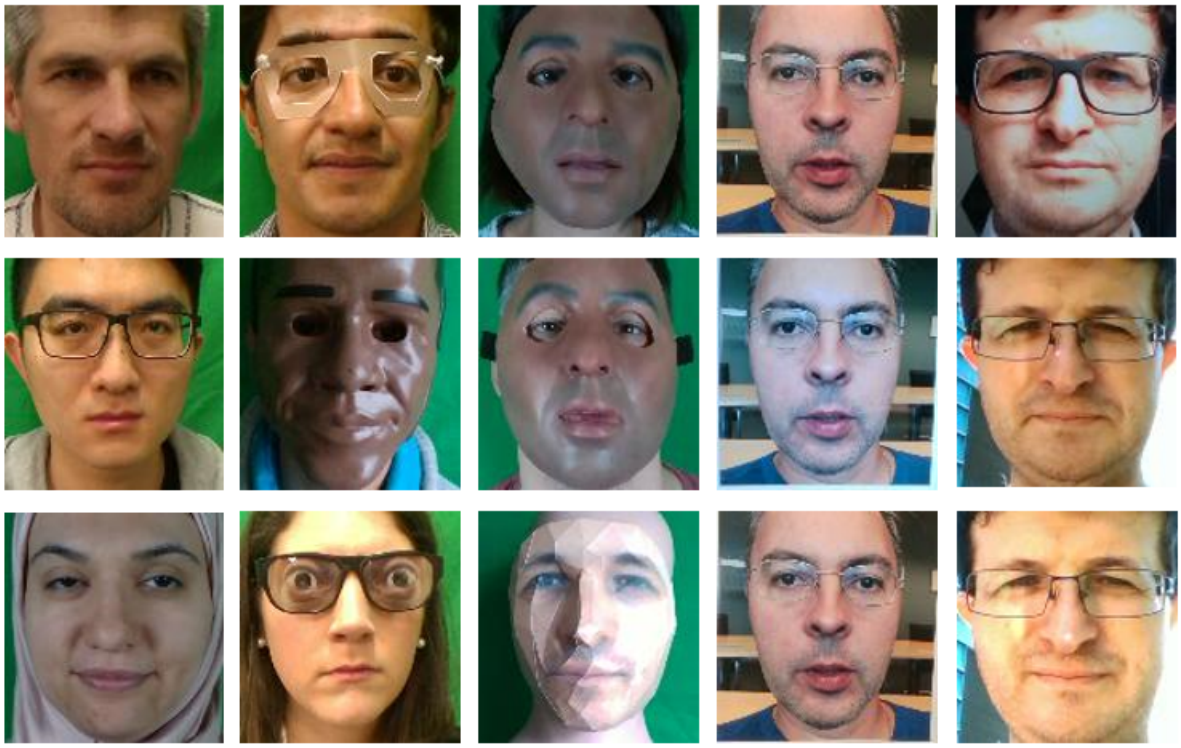


Figure 1.1: Examples of different PA types that a face biometric system may encounter.

It is evident from Figure 1.1 that some of the PA instruments would not easily confuse a human observer, but automatic facial recognition systems (without PAD)

are more susceptible to deception. Especially considering that there are two primary objectives of PA, to avoid identification, and to spoof another's identity.

1.2 Multi-Spectral Imaging

Visible light based cameras are ubiquitous, with components proliferating in consumer electronics, commercial security systems, law enforcement technology, and military surveillance. Visible cameras convert energy from reflected light in the visible spectrum (400 to 700 nm) to voltage levels that non-linearly map to pixel intensities. However, cameras can be designed to harness different regions of the electro-magnetic (EM) spectrum and thus capture different information about the subject. (examples: thermal, near infrared cameras, radio wave based imaging in astronomy)

Research in heterogeneous face recognition, specifically visible to near infrared (NIR) and visible to thermal, must overcome differences in quality, resolution, texture, and geometry between these imaging domains [Hu et al. (2017)]. The Structural Similarity (SSIM) Index was developed by Wang et al. (2004) to measure the "degradation" or structural differences between two images of the same scene. SSIM evaluates differences in local image structure rather than only using pixel-wise comparisons. Three main components for evaluation are used, luminance, contrast, and structure, each based on local means and variances.

Given two different images, x and y, SSIM(x,y) is defined as:

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_x y + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{1.1}$$

where μ_x, and μ_y are the local mean values and σ_x, σ_y, and σ_{xy} are the local standard deviations. C_1 and C_2 are constants that are included in the calculation of the luminance and contrast comparisons to add stability when the squared terms in the

denominator approach zero. C_1 and C_2 are determined by

$$C_1 = (K_1 L)^2, \tag{1.2}$$

$$C_2 = (K - 2L)^2 \tag{1.3}$$

where L is the dynamic range of the image and K_1 and K_2 are both small constants $K_1, K_2 \ll 1$.

Klare and Jain (2010) and Klare and Jain (2013) have shown that NIR to visible face recognition can be performed through linear discriminant analysis (LDA) applied on a collection of random subspaces. Results showed that shared or"common" discriminative projections can be learned such that NIR and visible images can be matched directly through these subspace projections.

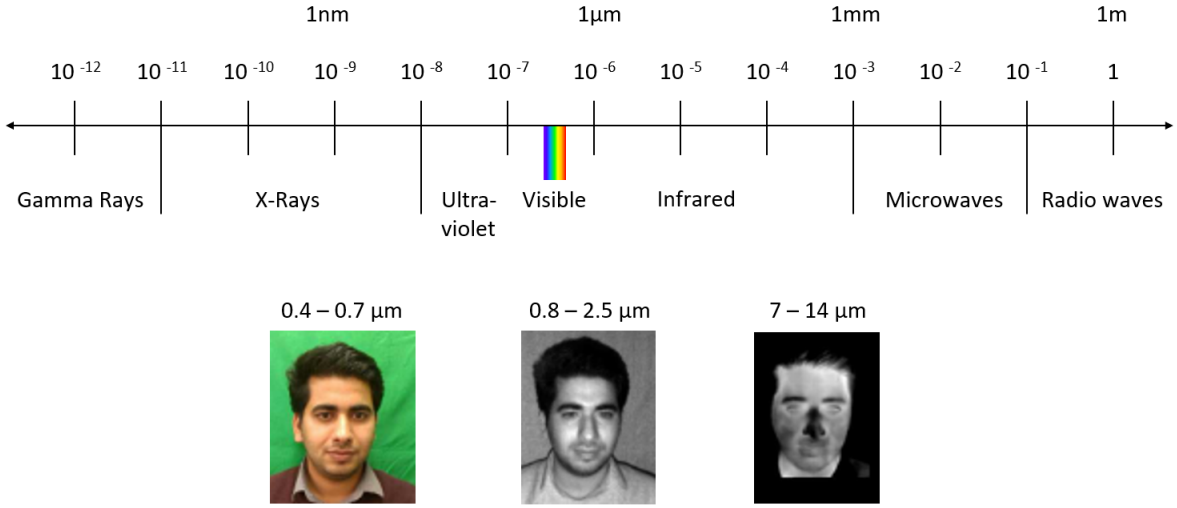
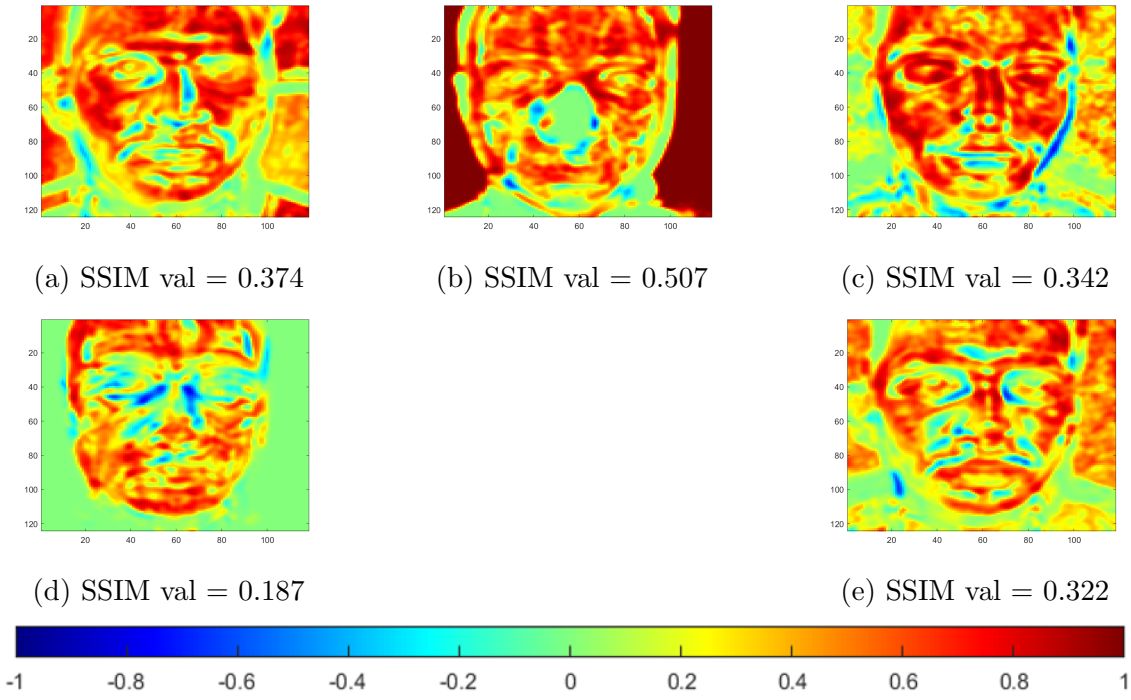


Figure 1.2: Imaging different regions of the electromagnetic spectrum provides different information on the subject.

In Figure 1.3 the domain shift from visible to infrared imagery in PAD data is shown using a sub-sample of the multi-spectral images used in this research. The SSIM value is measured between source and target images for the visible to thermal



(a) SSIM val = 0.374 (b) SSIM val = 0.507 (c) SSIM val = 0.342

(d) SSIM val = 0.187 (e) SSIM val = 0.322

(f) Colorbar

Figure 1.3: SSIM map comparisons where same domain pairs are compared with cross domain pairs of bonafide face images. (a) Visible to Visible (b) Thermal to Thermal (c) NIR to NIR (d) Visible to Thermal (e) Visible to NIR

and visible to NIR scenarios. By taking 100 images from each of the presentation attack sub categories in the WMCA dataset, we also investigate how the domain gap is effected by the type of attack. The average SSIM value for visible-thermal and visible-NIR pairs are shown in Table 1.1 for each category.

Both Figure 1.3 and Table 1.1 indicate that generally speaking there is higher similarity, or a smaller domain gap, between visible and NIR images. This agrees with the findings in Hu et al. (2017) where bonafide faces were evaluated across multiple infrared imaging domains.

Table 1.1: Average SSIM values are computed for each of the presentation sub categories.

Presentation	Thermal	NIR
Bonafide	0.183	**0.374**
Facial Disguise	0.155	**0.303**
Fake Face	0.187	**0.306**
Print	0.070	**0.204**
Video	0.086	**0.187**

1.2.1 Near Infrared

Near Infrared (NIR) light inhabits the part of the EM spectrum just beyond visible light, in the range of 800 to 2,500 nm. Like visible imaging, NIR imaging is still reliant on reflected energy. An NIR imaging system requires a light source that emits in the NIR spectrum as well as specialized filters that only transmit within a narrowly defined range.

Several works have used NIR as well as shortwave infrared (SWIR) as the primary media for the PAD problem. Steiner et al. (2016) used a SWIR images to perform "skin detection" as an anti-spoofing pre-processing step for an FR system. Heusch et al. (2020) uses the shortwave infrared channel (SWIR) of the HQ-WMCA dataset for PAD. Raghavendra et al. (2017) fused 7 different spectral bands within visible and NIR wavelengths for pad.

1.2.2 Thermal

Thermal imaging sensors, such as cooled thermal imagers or microbolometers, are often sensitive to radiation from mid-wave infrared (MWIR) or long-wave infrared regions of the EM spectrum; 3-5 micron or 7-14 micron respectively.

Thermal images capture information based on the temperature of an object or scene. Everything warmer than 0 Kelvin radiates thermal energy and the spectral

emissions are described, in idealized form, by black body radiation. Planck's law characterizes black body radiation with the following equation,

$$B_\nu(\nu, T) = \frac{2h\nu^3}{c^2} \frac{1}{e^{h\nu/kT} - 1}, \tag{1.4}$$

where B_ν is the spectral radiance for a given frequency and temperature, ν is the frequency of the electromagnetic radiation, T is the temperature of the black-body, h is Planck's constant, and k is the Boltzmann constant. Unlike visible images, thermal images are acquired using focal plane arrays composed to narrow band gap semiconductors, e.g. indium antimonide (InSb), lead selenide (PbSe), or mercury cadmium telluride (HgCdTe). These semiconductors convert low energy (0.05-1.3 eV) from infrared photons to electric voltages which are quantized to discrete pixel intensity values representing relative (or absolute) surface temperatures. Since thermal focal plane arrays are sensitive to thermal emissions, the imagery produced lacks the high frequency detail and texture information present in reflection-dominant (e.g. visible) imagery.

Images of human faces captured by thermal cameras maintain some basic structures in common with visible face imagery. Boundaries between different regions are distinguished by changes in temperature from one area to the next. Thermal images depend a lot more on the content of an object than just the surface appearance, and regions with high capillary density will appear brighter (warmer) than areas with less blood flow or that protrude farther from the body (ears, nose). Any areas covered by hair such as the scalp or chin under a beard appear darker since hair has no mechanism for producing body heat.

However, a bonafide thermal face signature is significantly more challenging to fake since simple spoofing mediums, such as print or video based methods, do not

exhibit heat signatures that resemble human faces.

For this reason, thermal imagery frequently increases the performance of PAD systems. In one of the first PAD frameworks to use thermal imagery, Dhamecha et al. (2013) employed a patch-based "biometric" vs. "non-biometric" classification system that used thermal data to help identify which regions of the face might be under disguise.

1.3 Multi-Spectral PAD

Despite the benefits of combining information from multi-modal cameras for PAD applications, the added cost and complexity of using multiple sensors (e.g., visible, NIR, thermal, and depth) severely limits the use of PAD to local controlled access environments. Therefore, to leverage current (and future) surveillance camera infrastructure that is mostly comprised of visible cameras, we aim to learn to extract discriminative information (e.g., infrared imagery) from visible imagery using new domain adaptation objectives.

Recently, George et al. (2020) introduced the Wide Multi-Channel presentation Attack (WMCA) dataset that contains both 2D and 3D presentation attacks with spatially and temporally synchronized imagery across four different sensor domains. The WMCA dataset contains eight different kinds of presentation attack instruments (PAIs) that fall under four main categories. These attack categories include facial disguise (plastic halloween masks, paper glasses, funny eye glasses), fake face (mannequin, flexible silicon masks, paper masks), photo (print/electronic images), and video.

Other multi-modal PAD datasets include: Casia-Surf(Zhang et al. (2020a)), MLFP (Agarwal et al. (2017)), Multispectral-Spoof (MSSpoof) (Chingovska et al. (2016)),

3DMAD (Erdogmus and Marcel (2013)), as well as Casia-Surf CeFa (Liu et al. (2021)). Many approaches utilize information from all available imaging modes in order to carry out the PAD task, which requires complex and expensive sensor suites to perform PAD.

In this work, we enhance the performance of PAD systems that utilize readily available visible spectrum cameras and equipment by harnessing the auxiliary information present in supplementary image domains during the training process. However, NIR cameras with filters and thermal cameras need not be present at deployment. Our primary contributions include:

1.4 Contributions and Thesis Organization

This thesis makes contributions to the problem of face PAD through domain adaptation principles. The proposed framework, including CD-PAD, IDR, and DDA, enhances visible-based PAD performance by learning to predict information from discriminative infrared imagery with visible imagery during development. The organization of the following chapters is presented below.

- Chapter 2 lays out the related work that has been done utilizing multi-spectral PAD imagery and domain adaptation techniques employed in computer vision applications.

- Chapter 3 formalizes the methods used in the proposed framework and provides network architecture used.

- Chapter 4 provides details regarding the datasets used in experimentation, ablation studies of architectural components, and both qualitative and quantitave analysis of experimental results.

- Chapter 5 includes further discussion into the methods used in this thesis.
- Chapter 6 provides discussion regarding limitations due to data quality.
- Chapter 7 explains the conclusions of the work and potential for further investigation.

Chapter 2

Background

PAD spans a variety of approaches that harness information from visible cameras as well as those that rely on input from multi-spectral arrays. This chapter explains the relevant existing research on different PAD methods starting with those that have been developed for visible input data and then expanding to multi-modal approaches that require data from multiple imaging domains. Lastly, we discuss the general problem of domain adaptation and several approaches that we apply to the PAD task.

2.1 Visible Spectrum PAD

This section covers prior research in PAD applied to visible imagery which includes approaches such as image quality/artifact analysis, liveness detection, and SVM and deep learning methods.

2.1.1 Quality-based Methods

Quality-based, or texture-based, methods encompass approaches to PAD that specialize in two-dimensional(2D) two dimensional attacks by focusing on image degradation. The unifying theme of these methods is that they recognize that both print attacks

and video replay attacks introduce subtle patterns that are not present in a bonafide face sample.

Local Binary Patterns (LBP) is a method originally designed for texture analysis by Ojala et al. (2002). LBP is an operator that assigns each pixel a value depending on the pattern of the neighboring pixels (default follows the outer ring of a 3x3 block surrounding the center pixel, but other radius sizes can be used). Histograms of the resulting scores were then used to classify the types of patterns seen (various textiles, scale patterns, animal skin, etc) showing the method to be rotationally invariant.

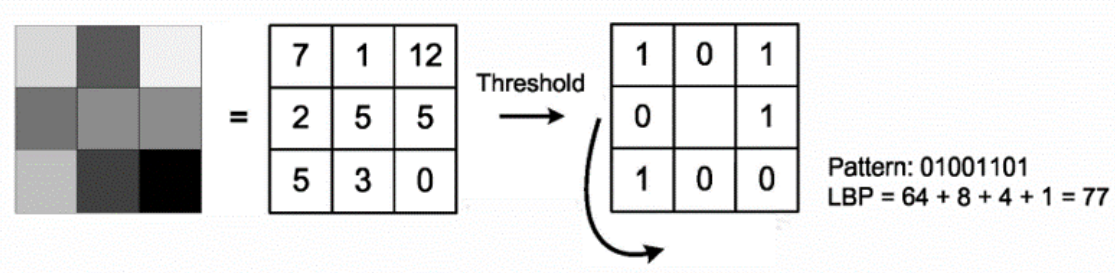


Figure 2.1: Example of the local binary pattern calculation process for a 3x3 region. Source Määttä et al. (2011)

LBP was recognized as a useful approach for 2D PAD applications due to the subtle patterns introduced by the attack media (print, video display) and has driven early research in texture analysis based approaches to PAD.

Inspired by the assumptions behind classic image de-noising techniques, that a degraded image can be resolved into the original image and some kind of additive noise, Jourabloo et al. (2018) developed a method that they call “Face De-spoofing”. Textures created by the spoofing medium (artifacts due to printing, screen quality) are treated as a specific type of noise. The spoofing noise was then modeled from a given attack sample while reconstructing the bonafide face from the spoof image. The spoof “noise” is used by the system to make a decision regarding the bonafide status of the presentation.

Other texture based methods methods (e.g., de Souza et al. (2017)) have used local binary patterns (LBP) to address 2D attacks, like print or replay PAs. LBP image representations have been primarily applied to visible imagery for PAD since infrared and depth imagery exhibit relatively fewer high frequency details (e.g., texture) compared with visible images.

One drawback to texture based methods is that they are primarily equipped to handle 2D attacks (i.e. print, video replay) and are not suitable for more subtle 3D attacks. Another issue faced by texture based methods is described by Agarwal et al. (2019) who show that "that simple intensity transforms such as Gamma correction, log transform, and brightness control can help an attacker to deceive face presentation attack detection algorithms." The Gamma corrections are defined by,

$$I_{out} = \alpha \cdot I_{in}^{\gamma} \tag{2.1}$$

where α is a constant set to 1, $\gamma = 0.5$, and I_{in} and I_{out} represent the intensity of a given input pixel and the intensity of the same pixel after Gamma correction. Log transformations enhances the darker areas of an image and is defined as,

$$I_{out} = c \cdot log(1 + I_i n) \tag{2.2}$$

where c is a constant defining the amount of the transformation applied to the image, $c = 2$, and 4 in this evaluation.

Atoum et al. (2017) introduced a dual CNN method utilizing randomly selected local regions from the face. One of the CNNs uses these image patches to generate individual scores for each patch that rate the likelihood that the given patch came from a spoof or bonafide image. Dividing the image into smaller pieces addresses the concern of overfitting caused by many PAD datasets containing a relatively small

number of subjects (compared to large scale face datasets used for FR). The second CNN is used to generate an estimation of the depth map for the input face image, such that bonafide presentations produce a face shaped map and attacks produce a flat depth map. This approach is suitable for detection of 2D attacks only since a mask will still produce a depth map resembling that of a human face. Several other methods Liu et al. (2018), Shao et al. (2019), have similarly used depth estimation as a regularization technique as one part of their PAD pipeline.

Shao et al. (2019) present a dataset domain generalizing approach that trains on three combined PAD datasets and is evaluated on a disjoint dataset. The proposed method used a mutli-adversarial framework to learn the generalized feature space. This approach also uses depth estimation as an auxiliary regularizing technique for PAD.

Zero-shot learning refers to the problem where a network learns from a set of examples of "known" classes and then learns to identify novel classes. Early applications of zero-shot learning to PAD by Arashloo et al. (2017), Xiong and AbdAlmageed (2018) involved only 2D attacks, where either print or video replay attack types would be present during training and the remaining attack type evaluated at test time. Liu et al. (2019) utilized a Deep Tree Network to perform zero-shot face anti-spoofing on a wider range of attacks that include several kinds of 3D attacks (i.e. multiple mask types, partial paper obfuscations).

2.1.2 Liveness detection

Liveness detection is a general approach to PAD based on the simple assumption that by detecting "signs of life" it can be determined whether or not a given presentation is real. These liveness cues have included: detection of a pulse, eye and head movements, etc.

The detection of heart rate using video frames is Remote Photoplethysmography (rPPG), a technique that hinges on the fact that light reflected from the skin will have different RGB color values depending on the blood flow beneath the surface. Liveness detection methods that use rPPG information extract patches from high arterial density (usually the forehead) for the best chances of detecting a signal.

Liu et al. (2018) use rPPG as a means of auxiliary supervision (in addition to depth estimation) for a deep learning PAD model. Figure 2.2 illustrates the combined rPPG and depth estimation approach.

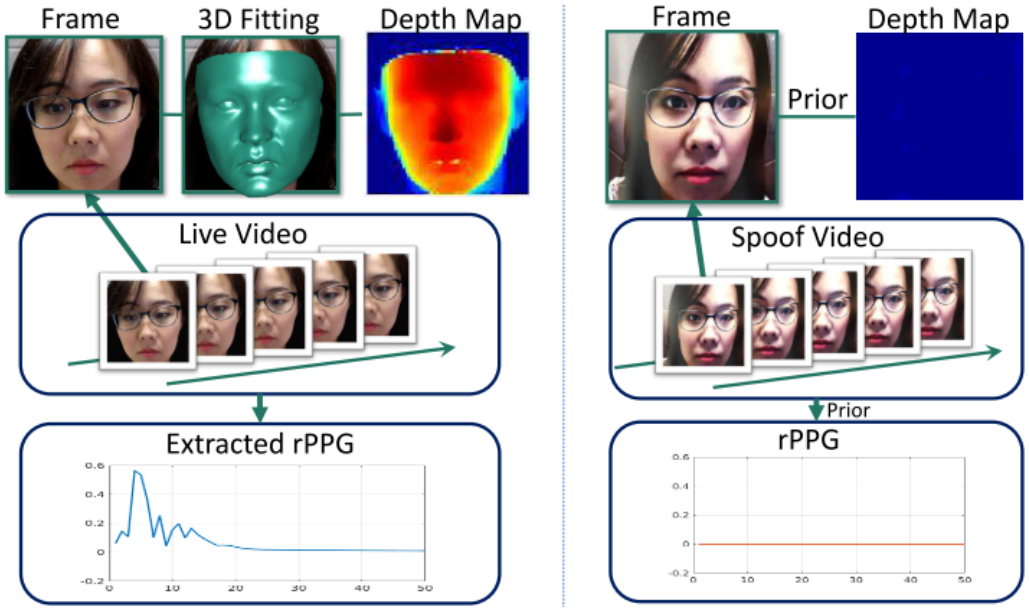


Figure 2.2: Bonafide vs. attack input and liveness signals. Source: Liu et al. (2018)

Heusch and Marcel (2018) apply the method of long-term spectral statistics (LTSS), first developed for speech based presentation attack detection, as a means of enhancing performance of face presentation attack detection through rPPG signals. LTSS is a general approach to signal processing, where the discrete Fourier Transform (DFT) is applied to an input signal, typically audio, to generate the first and second order statistics of the frequency components of the signal. For this study, Heusch and

Marcel (2018) use three different algorithms to generate the rPPG signals which are then processed using LTSS. They found that using LTSS in an rPPG based PAD pipeline improves over other rPPG based methods. The approach has the highest performance out of rPPG based methods, however it does not meet the state of the art performance for the datasets used. The authors of the study believe that rPPG is a promising method against subtle unseen 3D attacks, but there is still progress to be made.

2.2 Fusion based methods

We define fusion methods as any approach to PAD that utilizes multiple sources of information (visible, depth, thermal, etc) to make the final decision regarding the bonafide status of the presentation. Fusion can occur at different levels within the deep learning pipeline (score level fusion, feature level fusion, data level fusion). Techniques that rely on fusion have the advantages of utilizing complementary information from different sensing domains, but have the downside of requiring additional sensors deployed in the system.

Jiang et al. (2019) aimed to harness the complementary information in both visible and near infrared imaging modalities by building a multi-level fusion network. Information is combined through data level fusion, concatenating visible and NIR images, feature level fusion, concatenating the feature vectors output of the CNN, and score level fusion, concatenating the scores based on the individual modality inputs. Three separate branches of the network, sometimes containing two data streams individually, are integrated for this fusion technique to create a model with potentially high complexity and computational costs. This research does provide useful insight into which fusion level provides the best benefit. An ablation study shows that the

results of the score level fusion network are only slightly less accurate than the full multi-fusion network indicating that while there is a net gain from providing the different levels, the other channels might not provide enough benefit to outweigh the cost.

Raghavendra et al. (2018) investigated a specific niche of PAD which the authors refer to as disguise detection. This is specifically the issue of a subject choosing to adorn their face in such a way that their identity is concealed without creating an appearance that is obviously unnatural (i.e. facemasks, headcoverings). The specific mode of presentation attack studied is the application of a realistic false beard. Data was collected over 8 narrow spectral bands that include both visible and NIR wavelengths (530 to 1000 nm). The approach is focused on selecting image regions of 11x11 pixels from the moustache region of the face to train a deeply coupled autoencoder to generate the spectral signatures of the image patches and ultimately produce a decision for the presentation. This approach has the limitation of being tailored to a very specific attack scenario that is not necessarily common for all FR systems and will not generalize to other attack types.

To quantify the threat to security systems that is actually posed by increasingly realistic 3D masks Bhattacharjee et al. (2018) created the XCSMAD dataset with the latest in 3D modeled silicon masks. Using 3D imaging technology, life-like silicon masks can be manufactured that capture more detail than the cheap, mass produced masks that had previously been studied for obfuscation attacks. Three popular face recognition models, VGG-Face, Light CNN, and FaceNet, were evaluated with the XCSMAD dataset. The end result shows that the FR systems in the study are more than 10 times as likely to match a silicon mask to a real identity as they are likely to match a bonafide sample to an incorrect identity. This indicates a vulnerability to more advanced 3D attacks.

Kotwal et al. (2019) investigate the problem of increasingly subtle advancements in the technology behind custom made 3D masks for impersonation attacks. In this paper, both score level fusion and feature level fusion are evaluated on multi-modal imagery that includes visible, NIR, high-resolution thermal, and low-resolution thermal images. For the first time, it was proposed to use a pre-trained FR network as a feature extractor for PAD without fine tuning any layers. Results of single mode ablation studies show that infrared range imaging provides more discriminable features than visible imagery for the proposed method as well as all of the baseline methods evaluated.

2.3 Multi-Channel Presentation Attack Detection

In George and Marcel (2021); George et al. (2020) the Multi-Channel Convolutional Neural Network (MCCNN) was introduced for PAD. First, George et al. (2020) proposed a multi-channel (i.e., multi-modal) fusion approach that combined information from four imaging modalities: visible, near infrared (NIR), longwave "thermal" infrared (LWIR), and depth to perform PAD using the MCCNN architecture. Then, in George and Marcel (2021) the MCCNN is used to address the concern of novel "unseen" attacks. For the same purpose, Zhang et al. (2020b) developed an autoencoder network that utilizes the WMCA dataset to perform anomaly-based spoof detection. The fundamental difference between these approaches and our work is that they exploit multi-modal imagery during inference. Instead, we exploit multi-modal imagery offline in order to enhance the discriminability of visible-based PAD.

2.4 Domain Adaptation

Within the field of deep learning, domain adaptation refers to the task of generalizing so that the information learned from a source dataset can be applied to a disjoint dataset with a different underlying distribution, or domain shift. For example, two datasets that contain the exact same classes (handwritten numbers, cats vs dogs, etc.) will have statistical differences based on the fact that data collection practices will differ slightly, lighting conditions change, and so on. This is a significant concern in areas of study that involve the face, such as facial recognition, where regional and ethnic differences can affect things like skin tone, facial structure, and hair style to the extent that it is important to ensure that a deep learning model will perform appropriately outside of the context on which it was trained.

Transfer learning is a simple approach to domain adaptation where a model that has been pre-trained on a, usually, large scale dataset is "fine tuned" using a portion of the target data Weiss et al. (2016). The model parameters do not need to be trained from scratch, which can be especially useful for applications where the target dataset is of limited size. This approach is used in Nikisins et al. (2019) where pre-training on RGB visible imagery creates a base model that is fine tuned on multi-channel PAD data.

There are several works that recognize that the domain differences between several visible based PAD datasets is not insignificant. These approaches focus on using domain adaptation techniques to improve dataset generalization for PAD tasks. In this area of PAD research, four benchmark PAD datasets are used to create four different evaluation scenarios, where three of the datasets are used in training and one is left out for evaluation (Wang et al. (2020)). To this end, Mohammadi et al. (2020) used a "feature divergence measure" based on the symmetric Kullback-Leibler divergence

of a given filter between domains A and B to address PAD dataset generalization.

Domain adaptation is also used to describe the process of bridging the gap between different types of data in which the domain gap is much more significant. Here we study the process of extracting infrared-like features from visible light based imagery. As expressed in Section 1.2, the differences between the infrared data and visible data are not trivial and require a nuanced approach to domain adaptation.

In the following sections several existing approaches to domain adaptation are introduced and will be evaluated in relation to the problem of cross domain PAD in Chapter 4.

2.4.1 Maximum Mean Discrepancy

The Maximum Mean Discrepancy (MMD) Gretton et al. (2007) is a measure that was proposed to evaluate the similarity between two distributions by computing distance between their reproducing kernel Hilbert space (RKHS) embeddings. MMD has been used as a metric for minimizing the distance between source and target domain representations Long et al. (2013); Rozantsev et al. (2019). A 3D CNN framework for PAD tasks was introduced in Li et al. (2018) that incorporated MMD regularization between dataset domains for improved generalization.

Let $S = \{s_1, ..., s_N\}$ and $T = \{t_1, ..., t_N\}$ be the sets of features of the source and target domains. In this particular problem each set has the same number of elements N, although in general that need not be the case. Then the squared MMD of S and T can be expressed as

$$\begin{aligned} MMD^2(S,T) = &\sum_{i,j}^{N} \frac{k(s_i, s_j)}{N^2} + \sum_{i,j}^{N} \frac{k(t_i, t_j)}{N^2} \\ &- 2\sum_{i,j}^{N} \frac{k(s_i, t_j)}{N^2}, \end{aligned} \tag{2.3}$$

where $k(\cdot, \cdot)$ is the kernel associated with the RKHS.

The main disadvantage of MMD is that there is no discrimination between bonafide and attack instances. Therefore, we investigate an alternative to MMD for domain adaptation in the context of PAD for facial recognition systems.

2.4.2 Siamese Networks for PAD

Siamese networks (Bromley et al. (1993)) have been used to tackle both domain adaptation problems Motiian et al. (2017) de Freitas Pereira et al. (2019) and PAD tasks Perez-Cabo et al. (2019). In de Freitas Pereira et al. (2019), a siamese network implementing contrastive loss is used for heterogenous face recognition between different imaging domains where images are mapped to a shared embedding space. Siamese networks work well when imaging domains are sufficiently close (see Lezama et al. (2017)). However, when imaging domains are further apart, they have been shown to under-perform. Moreover, siamese networks, which are trained to ideally perform well on multiple domains at the same time, end up performing sub-optimally in all domains. Instead, we focus on modeling the complex interrelationships between two domains for PAD.

2.4.3 Domain Invariance Loss

The Domain Invariance Loss (DIL) Fondje et al. (2020); Poster et al. (2021) is a regularization technique proposed for domain adaptation for thermal-to-visible facial recognition tasks. DIL uses a domain classification network that learns the probability that the features produced from an image belong to either the visible (P_{vis}) or thermal domain (P_{therm}). Since the ultimate goal is similarity between the visible and thermal representations, the domain classifier is trained such that the two distributions are indistinguishable from each other. Specifically, the domain classification labels are

constant, i.e., $P_{vis} = P_{therm} = 0.5$. The potential disadvantage of this approach is that the labels are always the same, which implies there is a risk of models never learning patterns associated with either domain. Therefore, we consider an alternative where such patterns are learned and used in a regularizing fashion.

Chapter 3

Cross-Domain Presentation Attack Detection Framework

The proposed PAD framework (Fig. 3.1) aims to enhance visible-based PAD using new "high-level" domain adaptation principles. First, we define the problem: cross-domain presentation attack detection (CD-PAD). Then, we introduce the core components of our framework:

1. a base network architecture–to extract discriminative image representations,
2. a new dense domain adaptation (DDA) subnetwork–to learn a mapping between visible (source) and infrared (target) imagery,
3. a new CD-PAD objective function–to encourage task-level (i.e., inference level) domain adaptation,
4. a new inverse domain regularization function– to disentangle spectral information (domain specific) from PA information.

3.1 Preliminaries

Enhancing PAD performance from visible spectrum imagery requires exploitation of subtle cues (e.g., specular reflections) to differentiate between bonafide faces and PAs.

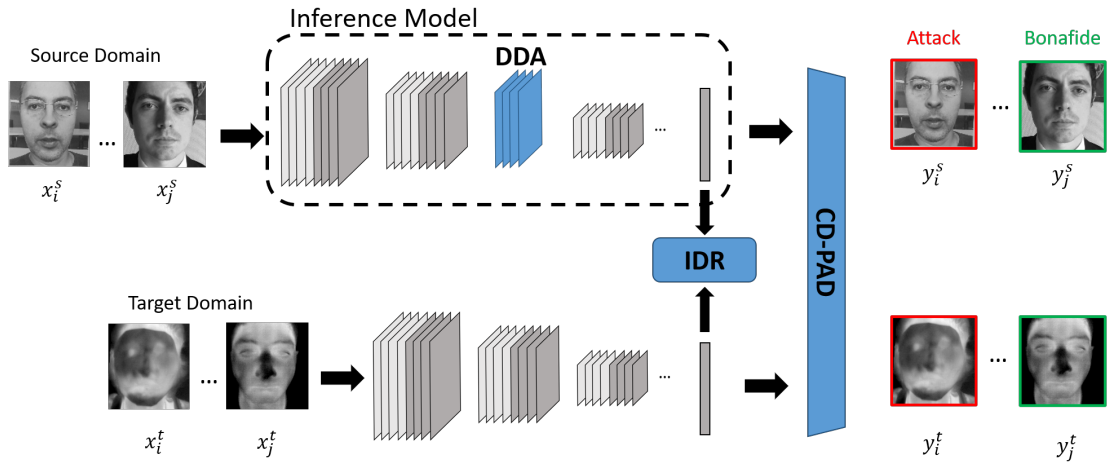


Figure 3.1: Schematic of CD-PAD with IDR regularization. The gray layers represent the convolutional layers that are not re-trained. All blue layers (DDA subnet, IDR, and CD-PAD classifier layers) are adapted during training. The DDA subnet is inserted into the inference model to learn the transformation of source imagery to the target embedding space. Bounding boxes at the output visually represent the final decision of the network.

To emphasize such subtle cues, we introduce a new CD-PAD framework. The CD-PAD problem is where discriminative information from a target domain is used to boost the quality of information extracted from the source domain. For example, by predicting infrared image representations from visible imagery, CD-PAD significantly improves the quality of visible-based PAD and reduces PAD system complexity (e.g. number/type of sensors) and cost.

Let $S = \{x_1^s, x_2^s, \ldots, x_n^s\}$ and $T = \{x_1^t, x_2^t, \ldots, x_m^t\}$ denote the sets of source (e.g., visible) and target (e.g., infrared) domain images, respectively. Here, n is the number of images from the source domain and m is the number of images from the target domain.

Let (x_i^s, x_j^t) denote a pair of source and target images with corresponding labels y_i^s and y_j^t. Unlike methods that use restrictive Euclidean distance metrics to bridge domain gaps, CD-PAD performs inference level domain adaptation which relaxes the

requirements for precise image registration/alignment and synchronous acquisition. Instead, the key requirement for CD-PAD is that both source and target labels sets, denoted by $\mathcal{Y}^s$ and $\mathcal{Y}^t$ respectively, must have overlapping labels. Mathematically, this requirement is

$$\mathcal{Y}^t \supseteq \mathcal{Y}^s, \tag{3.1}$$

where $y^t \in \mathcal{Y}^t$ and $y^s \in \mathcal{Y}^s$.

The main goal under our proposed CD-PAD framework is to learn a target domain PAD classifier, $P(y^t|f_t(x_j^t))$ that is sufficiently discriminative when used with source domain data, i.e., $P(y^t|f_s(x_i^s))$, where f_t is the mapping from the target domain to the associated latent subspace and f_s maps source imagery to the same "target" latent subspace. The primary objective for CD-PAD is to find an optimal source-to-target mapping f_s, such that $f_s(x_i^s) \approx f_t(x_j^t)$.

3.2 Base Architecture

George et al. (2020) showed that additional spectral data can increase the discriminative power of multi-modal PAD systems. However, many extant security systems employ visible spectrum cameras and use visible enrollment face imagery. Therefore, we propose a method that consists of training a PAD network to extract discriminative (e.g, infrared) representations from visible imagery while leveraging non-visible information only during training. The network contains two nearly identical data streams—one for processing source imagery and one for target imagery—consisting of CNNs with architectures based on the Light CNN network (Wu et al. (2018)). Both streams are fed into the proposed CD-PAD classifier. The primary difference is that the source stream is modified to include the addition of the DDA subnetwork, which is described in section 3.3.

Table 3.1: Light CNN Architecture for 124×118 pixel image

Layer	Filter Size /Stride	Output Shape $(H \times W \times C)$	Params
Conv1	$5 \times 5/1$	$124 \times 118 \times 96$	2,496
MFM1	—	$124 \times 118 \times 48$	–
Pool1	$2 \times 2/2$	$62 \times 59 \times 48$	–
Resblock1	$\begin{bmatrix} 3 \times 3/1 \\ 3 \times 3/1 \end{bmatrix} \times 1$	$62 \times 59 \times 48$	83,136
Conv2a	$1 \times 1/1$	$62 \times 59 \times 96$	4,704
MFM2a	—	$62 \times 59 \times 48$	–
Conv2	$3 \times 3/1$	$62 \times 59 \times 192$	83,136
MFM2	—	$62 \times 59 \times 96$	–
Pool2	$2 \times 2/2$	$31 \times 30 \times 96$	–
Resblock2	$\begin{bmatrix} 3 \times 3/1 \\ 3 \times 3/1 \end{bmatrix} \times 2$	$31 \times 30 \times 96$	332,160
Conv3a	$1 \times 1/1$	$31 \times 30 \times 192$	18,624
MFM3a	—	$31 \times 30 \times 96$	–
Conv3	$3 \times 3/1$	$31 \times 30 \times 384$	332,160
MFM3	—	$31 \times 30 \times 192$	–
Pool3	$2 \times 2/2$	$16 \times 15 \times 192$	–
Resblock3	$\begin{bmatrix} 3 \times 3/1 \\ 3 \times 3/1 \end{bmatrix} \times 3$	$16 \times 15 \times 192$	1,327,872
Conv4a	$1 \times 1/1$	$16 \times 15 \times 384$	74,112
MFM4a	—	$16 \times 15 \times 192$	–
Conv4	$3 \times 3/1$	$16 \times 15 \times 256$	442,624
MFM4	—	$16 \times 15 \times 128$	–
Resblock4	$\begin{bmatrix} 3 \times 3/1 \\ 3 \times 3/1 \end{bmatrix} \times 4$	$16 \times 15 \times 128$	590,336
Conv5a	$1 \times 1/1$	$16 \times 15 \times 256$	33,024
MFM5a	—	$16 \times 15 \times 128$	–
Conv5	$3 \times 3/1$	$16 \times 15 \times 256$	295,168
MFM5	—	$16 \times 15 \times 128$	–
Pool4	$2 \times 2/2$	$8 \times 8 \times 128$	–
Linear	—	512	4,194,816
MFM6	—	256	–

The Light CNN is a deep learning model that was designed to tackle face recognition, while also being robust to the issue of noisy labels in large face datasets and maintaining a small footprint with respect to storage requirements and computational

complexity. A major advantage of Light CNN is the application of the Max Feature Map (MFM) activation function. The MFM layer implements a type of neural inhibition by taking two feature maps and only passing the element-wise maximum values to the next layer. The MFM operating on two input feature maps can be expressed mathematically as,

$$\hat{x_{ij}^k} = max(x_{ij}^k, x_{ij}^{k+N}). \tag{3.2}$$

Figure 3.2 shows the MFM process. Several works including those by, Kotwal et al. (2019), George et al. (2020), and Kotwal et al. (2020) have found that Light CNN makes a favorable feature extractor for PAD when the model is pre-trained for face recognition.

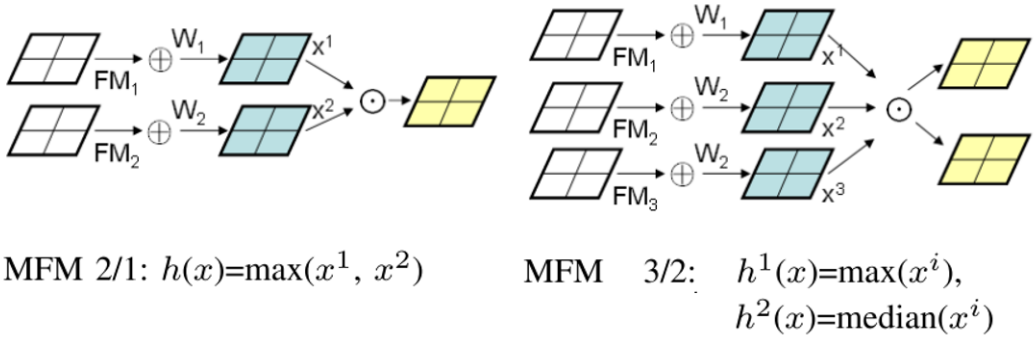


Figure 3.2: Shown are the two types of MFM used in Light CNN. Left: MFM 1/2 favors the strongest out of two neuron activations. Right: MFM 2/3 only suppresses one neuron out of three and keeps the two highest values. Source: Wu et al. (2018)

We use the Light CNN weights that have been pre-trained on the MSCeleb-1M dataset (Wu et al. (2018)), which is a large-scale face dataset containing 10 million images, as the feature extractor for both streams of the CD-PAD network. Then, transfer learning is applied to both streams to re-use relevant model parameters. This approach is similar to that used by George et al. (2020), except that our two-stream network is trained in a domain adaptive manner instead of a multi-modal fusion manner, meaning that only visible imagery is required during deployment opposed to

requiring both visible and infrared imagery. Table 3.1 summarizes each layer of the Light CNN architecture, which is comprised of convolution (Conv), Max-Feature-Map (MFM), max pooling, and residual layers.

3.3 DDA Sub-network

We select a developed architecture that would best enable the network to learn the mapping from source to target domain considering the constraints of the problem. In the simplest architecture, one layer connects directly to the next such that information flows in the following manner. The input and output of a given layer can be represented by the equations

$$y_i = A(x_{i-1}), \tag{3.3}$$

$$x_{i-1} = W_{i-1} \cdot x_{i-1} \tag{3.4}$$

where y is the post-activation for a given layer and x is the pre-activation, $A\dot{()}$ is the nonlinear activation function of the layer, and W_{i-1} represents the layer weights. This works just fine for small networks. However, the problem of vanishing gradient arises as networks grow deeper.

We explore two different connection types for cross domain PAD.

Dense connection – each layer l is connected to all previous $(L - l)$ layers.

$y_i = A([x_0, x_1, \ldots, x_{i-1}])$

Residual connection – The activation of layer l is summed with its input.

$y_i = A(x_{i-1}) + x_{i-1}$

Dense connections in neural networks were introduced by Huang et al. (2017) to address the vanishing gradient issue by giving each layer a direct connection to

the objective function. Due to the information sharing from concatenating all of the feature maps from previous layers, densely connected networks can perform well with fewer parameters than other leading architectures.

He et al. (2016) took a different approach to the problem of vanishing gradients and degradation faced by deep networks. The authors restrict the layers of their network to explicitly fit a residual mapping $\mathcal{F}(x) + x$. These networks employ skip connections to build residual blocks, shown in Figure 3.3.

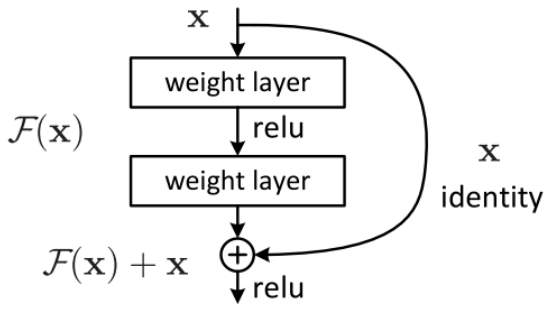


Figure 3.3: The basic building block of a Residual Network. Source: He et al. (2016)

Table 3.2: DDA Subnetwork Architecture for 124×118 pixel image

Layer	Inputs	Output Shape	Params
$BatchNorm$	Pool3	$16 \times 15 \times 192$	384
δ_{conv1}	$BatchNorm$	$16 \times 15 \times 48$	82,992
δ_{conv2}	δ_{conv1}	$16 \times 15 \times 48$	20,784
δ_{conv3}	$[\delta_{conv1}, \delta_{conv2}]$	$16 \times 15 \times 48$	41,520
δ_{conv4}	$[\delta_{conv1}, \delta_{conv2}, \delta_{conv3}]$	$16 \times 15 \times 48$	62,256
δ	$[\delta_{conv1}, \delta_{conv2}, \delta_{conv3}, \delta_{conv4}]$	$16 \times 15 \times 192$	—

A domain adaptive subnetwork is added to the visible (source) stream of the CD-PAD network to learn the mapping from the source to target domain. We propose a new Dense Domain Adapation (DDA) subnetwork which is composed of a dense block Huang et al. (2017) that consists of four convolutional layers as shown in Table

3.2. Mathematically, the DDA subnetwork is represented as

$$\delta(u) = Concat\{\delta_{conv1}(u), \delta_{conv2}(u), \delta_{conv3(u)}, \delta_{conv4}(u)\}, \tag{3.5}$$

where

$$\delta_{conv1}(u) = ReLU(Conv(BatchNorm(u))), \tag{3.6}$$

$$\delta_{conv2}(u) = ReLU(Conv(\delta_{conv1}(u))), \tag{3.7}$$

$$\delta_{conv3}(u) = ReLU(Conv([\delta_{conv2}(u), \delta_{conv1}(u)])), \tag{3.8}$$

$$\delta_{conv4}(u) = ReLU(Conv([\delta_{conv3}(u), \delta_{conv2}(u), \delta_{conv1}(u)])), \tag{3.9}$$

with $Conv(\cdot)$ respresenting a 3×3 convolution and $ReLU(\cdot)$ the rectified linear unit activation function. The parameters of the DDA subnetwork are optimized using our proposed CD-PAD loss (section 3.4).

The DDA subnetwork (Eq. 3.5) is motivated by the Residual Spectrum Transform (RST) subnetwork used by Fondje et al. (2020) who introduced a residual transformation [He et al. (2016)] based subnetwork to bridge domain gaps for thermal-to-visible face recognition. The effects of subnetwork type (i.e., residual versus dense) and placement within the Light CNN on the overall performance of CD-PAD are described in Section 4.6.2. The dense architecture was selected for the DDA subnetwork primarily due to superior performance observed in the context of cross domain PAD.

The DDA subnetwork receives the output of the Pool3 max pooling layer shown in the Light CNN architecture in Table 3.1 as input to the BatchNorm layer. The dense output of DDA is then passed to the Resblock3 layer of Light CNN and through the remainder of the network. Figure 3.4 illustrates the connections between the layers

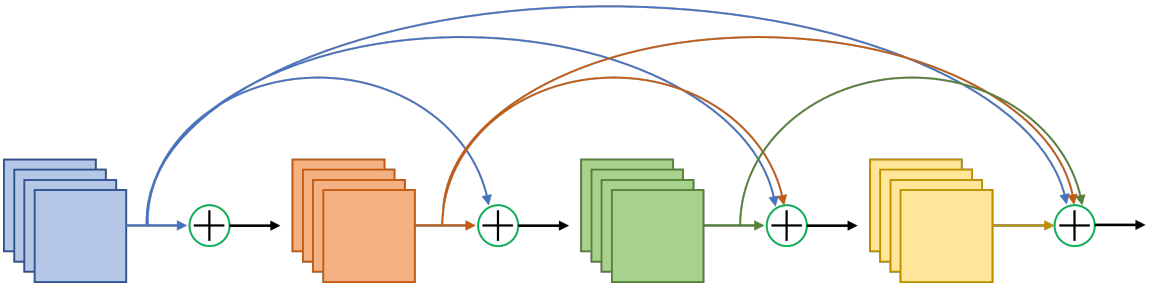


Figure 3.4: Diagram showing the connections between densely connected layers of the DDA subnetwork

of the DDA subnetwork.

3.4 Cross Domain Presentation Attack Detection

The proposed CD-PAD framework alternates training between the source and target domains to optimize information extracted from target domain face imagery to guide the adaptation of source domain representations. First, the PAD classifier is trained exclusively on the target data. The target domain classifier and Light CNN are trained in a manner to avoid over-fitting to the target data. We found that over-training on target data can lead to under-performing on source imagery (i.e., visible based PAD). Results from detailed source-target trade off ablation studies (shown in section 4.6.1) indicate when the target learning phase of training is complete. The PAD classifier weights are optimized by minimizing the Binary Cross Entropy (BCE) loss function between the labels and predictions by,

$$\begin{aligned}\mathcal{L}(x^t, y^t) =&(1 - y^t)\log(1 - f(x^t; w^t)) \\ &+ y^t \log(f(x^t; w^t)),\end{aligned} \tag{3.10}$$

where x^t and y^t are the source and target input images and labels from Section 3.1 and w^t refers to the classifier weights that are trained using the target imagery.

After this initial phase, the trained classifier optimizes the parameters of the DDA subnetwork that transform the source domain image representations to representations to exhibit properties similar to target domain imagery. In the domain adaptive phase, the classifier weights, w^t, remain fixed so that the objective function can only be minimized by transforming the feature representation of the visible domain. For the domain adaptive training, the BCE loss function is

$$\begin{aligned}\mathcal{L}(x^s, y^s) =&(1 - y^s)\log(1 - f(x^s; w^t)) \\ &+ y^s \log(f(x^s; w^t)),\end{aligned} \tag{3.11}$$

where w^t represents the classifier parameter weights that had previously been trained on the target data.

The CD-PAD framework ultimately works due to the fundamental assumption in Eq. 3.1, where we assume that both target and source domain span the same label sets. Due to the asynchronous, alternating training strategy used by CD-PAD, target and source imagery are not required to be precisely synchronized or co-registered. Therefore, CD-PAD is more flexible and extensible than existing domain adaptation methodologies, especially those that optimize Euclidean distances between corresponding pairs or triplets.

After training the CD-PAD framework, only the source stream (i.e., inference model in Fig. 3.1) is used for deployment of the PAD system. This provides a very efficient and cost effective solution for PAD.

3.5 Inverse Domain Regularization

Lastly, we propose a new inverse domain regularization(IDR) technique that aims to help guide the transformation of source domain representations to the target domain

subspace. Unlike Fondje et al. (2020) who used domain invariance loss to force a matching domain classification for both domains, we train a domain classifier to correctly differentiate between source and target domains. After learning the distinction, inverting the labels of the source data is what drives the domain adaptation provided by IDR.

First, the IDR domain classifier is trained with correct domain labels for each of the input images and learns to appropriately discriminate between the two domains. Using the same notation from Section 3.1, let $P^t(x_i)$ be the probability that a given training image x_i comes from the target domain. The IDR classifier is trained to predict $P^t(x_i^s) = 0$ and $P^t(x_i^t) = 1$, a correct classification of the feature domains for each of the input images. Next, in order to guide the network to map the source images to the target feature space, we implement the domain inversion of IDR. In this domain adaptive stage of training, the domain classifier parameters are fixed while the DDA subnetwork is updated. Here, the labels for the source images are intentionally labeled incorrectly as target images. The DDA subnetwork in the source channel must then adapt to transform the source features so that they will be identified as the target class by the domain classifier, or mathematically, that $P^t(x_i^s) = 1$. The bottom line is that IDR aims to reduce differences between source and target image representations in a class agnostic manner and thus complements the CD-PAD loss by imposing additional constraints.

Chapter 4

Experiments and Results

In this Chapter we describe the datasets used for evaluation of the CD-PAD method as well as the evaluation metrics that measure performance. Both qualitative and quantitative results for the full CD-PAD framework are presented in Sections 4.4 and 4.5. Lastly, multiple ablation studies evaluating the effects of different components of the CD-PAD framework are included in Section 4.6.

4.1 Datasets

To train any deep learning model, obtaining a dataset of sufficient size and quality is important. The best Presentation Attack datasets contain several modes of attacks, offering the ability to predict how a network will perform on an unknown or "unseen" attack type. For this project we focus only on PAD datasets that contain both visible and infrared imagery. Information about the datasets used for analysis is provided in Sections 4.1.1 through **??**.

4.1.1 WMCA

For training and evaluation on the WMCA dataset, the "grandtest" protocol referred to by George et al. (2020) is used. The data is split into three subsets: train, dev,

and test. For each domain, the subsets contain 28,223, 27,850, and 27,740 images respectively. The distribution of attack categories are consistent across each of the sets and individual subjects do not appear in multiple subsets. The test subset contains 5,750 bonafide images, 1,649 facial disguise images, 13,041 fake face images, 4,200 photo attack images, and 3,100 video attack images.

4.1.2 MSSpoof

To show CD-PAD's potential for generalization, we also evaluate on the MSSpoof (Chingovska et al. (2016)) dataset. MSSpoof contains both visible and NIR imagery of 21 individuals. Like WMCA, MSSpoof is split into three identity disjoint subsets: train, dev, and test. All of the PAs in the MSSpoof dataset are print style attacks. The training subset contains 594 visible images and 577 NIR images, the dev subset contains 398 visible images and 395 NIR images, and the test subset contains 396 visible images and 395 NIR images.

4.1.3 CASIA-SURF

CASIA-SURF is a largescale PAD dataset that contains three different imaging modalities: depth, NIR, and visible. All the attack types are variations of the standard print attack where a printed photograph is presented to the FR system. The print attacks in CASIA-SURF can be split in to 3 categories: photos with eye regions cut out, photos with eyes and nose cut out, and photos with eyes, nose, and mouth cut out. The dataset contains 1000 unique subjects and a total of 492,522 images.

For evaluation, the data is split so that the training subset contains bonafide faces and only half of the attack types (curved print with eyes/mouth/nose cut, flat print with eye/mouth/nose cut, and curved print with eyes/nose cut) so that the test

and validation subsets contain unseen attacks. Additionally there are 2 evaluation protocols for both within modal and cross modal evaluation

4.2 Implementation

All models were trained in PyTorch (Paszke et al. (2019)) and updated using the ADAM optimizer (Kingma and Ba (2015)) with a learning rate of $1x10^{-4}$. Features were generated using the Light CNN (Wu et al. (2018)) framework initialized with weights pre-trained for facial recognition. During the first phase of training, the fully-connected PAD classification layers are trained on thermal data. In the final cross domain training stage the weights of the DDA subnetwork in the visible data stream are made trainable. The second stage of training uses the same optimizer and learning rate. Networks trained with inverse domain regularization required an additional stage of training, wherein only layers in the parallel domain classification network are updated. Data augmentation is utilized during training with random horizontal flipping with a probability of 0.5, and random rotation of maximum 10 degrees.

Preprocessing on the MSSpoof dataset included 5-point facial landmark registration and tight cropping around the face. Image cropping is utilized to alleviate potential problems with over-fitting as a result of the limited quantity of data in MSSpoof. Restricting the network to only learn from information contained in the face prevents it from focusing on background details that are often dataset specific.

4.3 Evaluation Metrics

4.3.1 ROC Analysis

The Reciever Operating Characteristic curve (ROC curve) is a tool used to analyze the performance of binary classification systems. The ROC curve is a plot of the True Positive Rate (TPR) vs. False Positive Rate (FPR) where the decision threshold τ is swept from the minimum to the maximum value of the decision scores that are under evaluation. In a binary classification system there are four different decision scenarios that relate to the ROC analysis. For input where the ground truth label is positive, the system may either return a positive or negative score producing True Positive (TP) or False Negative (FN) decisions respectively. Alternately a sample might have a ground truth label of negative. Then a positive score would give a False Positive (FP) decision and a negative score would give a True Negative (TN) decision. The FPR and TPR are given by the equations,

$$TPR(\tau) = \frac{TP(\tau)}{TP(\tau) + FN(\tau)}, \tag{4.1}$$

$$FPR(\tau) = \frac{FP(\tau)}{FN(\tau) + TN(\tau)}. \tag{4.2}$$

where τ is the decision threshold.

4.3.2 ACER Metrics

Results are reported according to the ISO/IEC 30107-3 standard metrics for presentation attack detection, Attack Presentation Classification Error Rate (APCER), Bonafide Presentation Classification Error Rate (BPCER), and the Average Classification Error Rate (ACER) ISO/IEC 30107-3:2017. APCER designates the proportion of presentation attacks incorrectly identified as bonafide presentations, and BPCER

is the proportion of incorrectly identified bonafide presentations. The metrics are defined as follows:

$$APCER(\tau) = \frac{FP(\tau)}{FP(\tau) + TN(\tau)}, \tag{4.3}$$

$$BPCER(\tau) = \frac{FN(\tau)}{FN(\tau) + TP(\tau)}, \tag{4.4}$$

$$ACER(\tau) = \frac{APCER(\tau) + BPCER(\tau)}{2}, \tag{4.5}$$

where FN, FP, TN, and TP are the number of false negatives, false positives, true negatives, and true positives for a given threshold τ as introduced in Sec. 4.3.1.

Results are reported by giving the BPCER value at selected APCER values of 1% and 5%. This provides a consistent operating point for comparison in the ROC curves and provides insight into the performance at both a low and slightly less strict false positive rate.

4.4 Qualitative Analysis

Quantitative metrics are important for providing a hard line of comparison between different methods, but do not always give the full picture behind the performance of a model. In this section we present qualitative analysis of the models that provide more visual context for the performance.

4.4.1 WMCA

To illustrate the enhancements due to CD-PAD, we evaluate the feature representations of the visible imagery (bonafide and PA samples) using the t-Distributed Stochastic Neighbor Embedding (t-SNE) (van der Maaten and Hinton (2008)) method

to visualize the data. The t-SNE representations for the visible baseline and the CD-PAD adapted visible features are shown in Fig. 4.1. Data samples used for the t-SNE visualization are randomly selected from the test set. The adapted features are generated from a CD-PAD + IDR network trained with thermal imagery as the target domain. It is clear from the plots that the cross domain adaptation causes the bonafide samples to be more tightly clustered in the feature space and have less overlap with the attack samples.

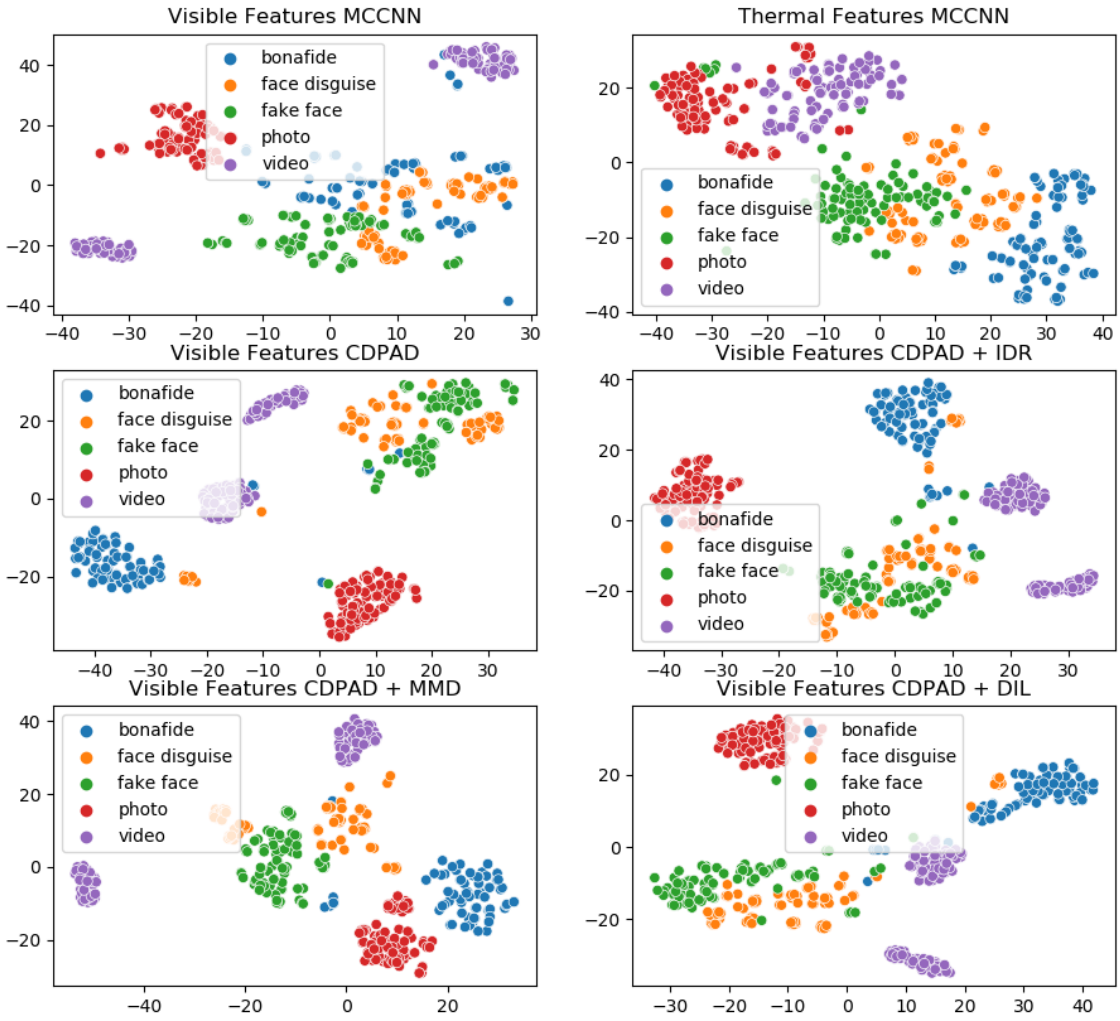


Figure 4.1: Compared to the single mode visible baseline, our method shows better separability between bonafide and all attack data points.

4.4.2 MSSPOOF

To illustrate the effects of the CD-PAD framework, analysis of the raw predictions for the MSSPOOF test set are shown. Since PAD is treated as a 2 class (binary) problem, the final layer generates a confidence score in the range of [0,1] that indicates whether or not the image is a bonafide presentation.

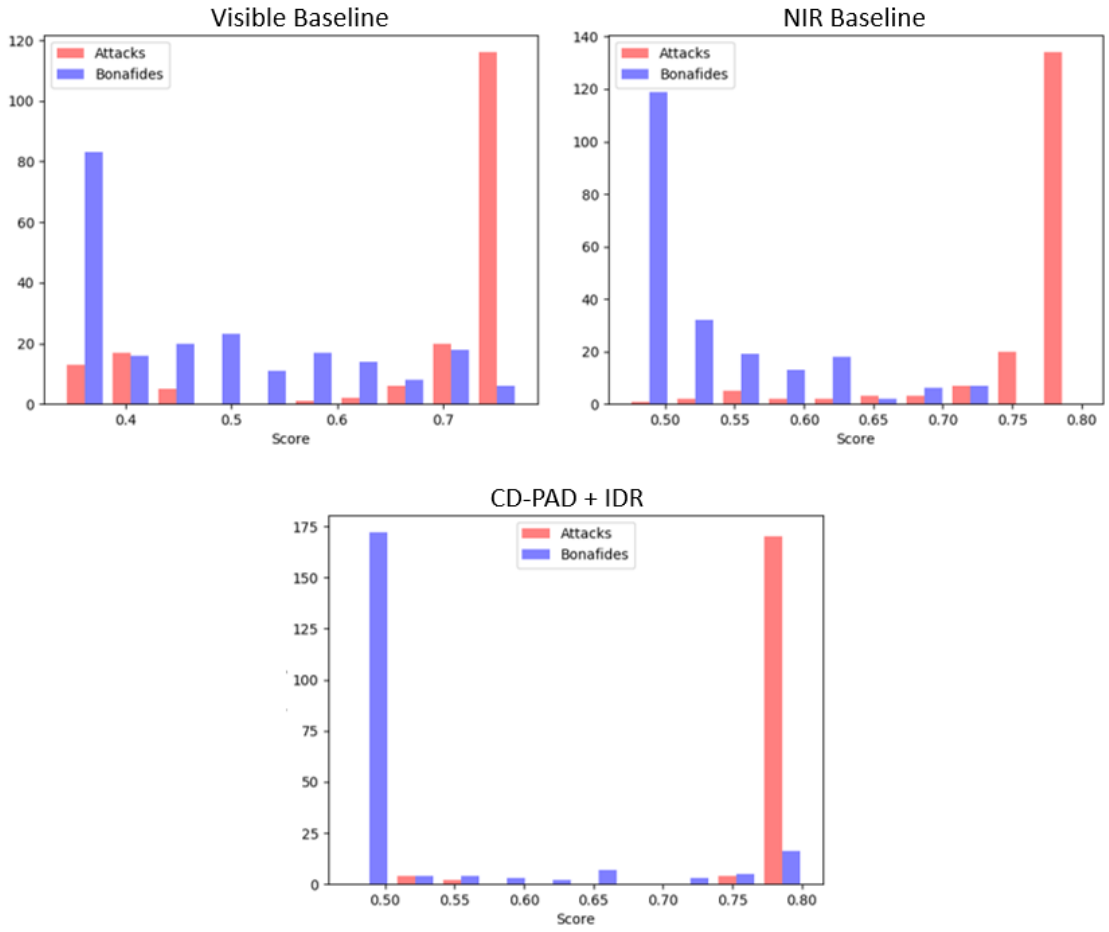


Figure 4.2: Histograms showing the distribution of scores

4.5 Quantitative Results

Next, we compare the performance of CD-PAD with visible and infrared (thermal or NIR) baseline models using WMCA and MSSpoof. The thermal and NIR specific models represent the upper performance bounds that can be attained by adapting visible data via our CD-PAD framework.

4.5.1 WMCA

We compare the results of the CD-PAD method using two different target domains, thermal and NIR, against networks trained for the PAD task on single modal data. The CD-PAD method improves upon the quality of the attack detector when only visible data is available in a deployment scenario.

Table 4.1: CD-PAD results where NIR is the target domain using the WMCA dataset

Method	BPCER @ 1% APCER	BPCER @ 5% APCER	AUC
MCCNN(NIR) George et al. (2020)	5.93 ± 6.54	1.54 ± 1.94	0.997 ± 0.003
MCCNN(Visible) George et al. (2020)	74.59 ± 9.87	43.72 ± 9.43	0.895 ± 0.029
Siamese network	26.08 ± 3.16	11.18 ± 1.97	0.957 ±0.008
CD-PAD	18.7 ± 1.77	9.95 ± 1.31	0.962 ± 0.008
CD-PAD+DIL	19.84 ± 0.34	11.2 ± 1.41	0.970 ± 0.002
CD-PAD+MMD	20.9 ± 4.1	13.1 ± 2.24	0.977 ± 0.001
CD-PAD+IDR	**17.13**± 1.38	**9.27** ± 2.13	**0.980** ± 0.000

The effects of using NIR imagery as the visible adaptation target are shown in Table 4.1. The CD-PAD network greatly improves over the visible baseline. With NIR as the target domain CD-PAD achieves an average of 18.7% BPCER at a 1% APCER operating point, improving over the visible baseline by 55.89%. Adding IDR to the CD-PAD framweork results in an additional improvement of 1.57%.

Table 4.2: CD-PAD results where thermal is the target domain using the WMCA dataset

Method	BPCER @ 1% APCER	BPCER @ 5% APCER	AUC
MCCNN(Thermal) George et al. (2020)	3.83 ± 2.45	0.0 ± 0.0	0.998 ± 0.001
MCCNN(Visible) George et al. (2020)	74.59 ± 9.87	43.72 ± 9.43	0.895 ± 0.029
Siamese network	36.18 ± 3.17	19.89 ± 6.86	0.939 ± 0.014
CD-PAD	24.3 ± 2.36	8.75 ± 1.64	0.973 ± 0.003
CD-PAD+DIL	19.31 ± 1.25	7.88 ± 2.83	0.981 ± 0.006
CD-PAD+MMD	48.24 ± 0.79	23.4 ± 1.98	0.948 ± 0.001
CD-PAD+IDR	**12.42** ± 0.52	**6.90** ± 1.56	**0.982** ± 0.007

Table 4.2 shows the results when thermal imagery is available for cross domain training. When CD-PAD is used on its own, the visible based PAD results are boosted. CD-PAD shows a marked improvement in the BPCER at low APCER operating points in the ROC curve. CD-PAD achieves an average of 24.3% BPCER at a 1% APCER operating point, and improves by 50.29% over the visible baseline. Including additional domain adaptation loss components had varying effects on the CD-PAD performance. Introducing MMD to help guide domain adaptation actually hurt performance. However, the combination of CD-PAD and IDR using the thermal target imagery achieved the biggest improvement in visible based PAD on WMCA decreasing the BPCER at a 1% APCER by 62.17%.

4.5.2 MSSpoof

In Table 4.3, we evaluate the CD-PAD method using the MSSpoof dataset where visible source imagery is adapted to the target NIR domain. Once again, CD-PAD improves upon training the model on visible imagery alone. The network trained on visible imagery achieves an average of 42.85% BPCER at a 1% APCER operating

point, and the CD-PAD network achieves 13.75% for the same metric. Adding IDR to the CD-PAD framework offers a small performance boost to the BPCER score at 1% APCER, improving the CD-PAD performance by 0.27%. The overall improvement due to CD-PAD + IDR

Table 4.3: CD-PAD results using MSSpoof. NIR is the target domain.

Method	BPCER @ 1% APCER	BPCER @ 5% APCER	AUC
MCCNN(NIR) George et al. (2020)	16.27 ± 3.74	12.09 ± 3.1	0.977 ± 0.003
MCCNN(Visible) George et al. (2020)	42.85 ± 0.54	27.67 ± 0.51	0.891 ± 0.025
CD-PAD	13.75 ± 0.59	9.25 ± 0.3	0.987 ± 0.001
CD-PAD+DIL	14.74 ± 2.39	**8.11** ± 0.81	0.987 ± 0.001
CD-PAD+MMD	28.36 ± 3.76	17.54 ± 0.26	0.976 ± 0.037
CD-PAD+IDR	**13.48** ± 2.02	10.46 ± 2.64	**0.987** ± 0.002

4.5.3 CASIA-Surf

Table 4.4: CD-PAD results on Casia-Surf, where NIR is the target domain.

Method	BPCER @ 1% APCER	BPCER @ 5% APCER	AUC
MCCNN(NIR) George et al. (2020)	85.98± 0.07	65.93 ± 3.78	0.843 ± 0.028
MCCNN(Visible) George et al. (2020)	65.64 ± 12.79	39.35 ± 15.89	0.911 ± 0.048
NIR (with DDA)	19.42 ± 9.93	3.64 ± 2.41	0.991 ± 0.004
Visible (with DDA)	63.16 ± 11.01	35.66 ± 13.95	0.923 ± 0.042
CD-PAD*	**55.22 ± 4.2**	**21.94 ± 2.4**	**0.957 ± 0.003**
CD-PAD*+DIL	58.40 ± 1.53	29.9 ± 0.45	0.944 ± 0.000
CD-PAD*+MMD	74.10 ± 1.36	33.87 ± 1.70	0.939 ± 0.002
CD-PAD*+IDR	78.7 ± 5.78	36.14 ± 5.61	0.929 ± 0.007

Table 4.4 shows the results of the CD-PAD method with adjustments made to accommodate changes in image quality in the CASIA-Surf dataset (CD-PAD* indi-

cates that the CD-PAD method involves two DDA subnetworks). Additional analysis into the motivations behind the changes made to CD-PAD is explained in Chapter 5. In the original single mode baselines, the network struggled to sufficiently learn from the target data in order to offer an improvement through cross domain training. For the experiment on CASIA-Surf, a DDA subnetwork is added to both streams of the CD-PAD network. These subnetworks learn different transformations and are not "shared" between the source and target streams. In this situation, the best improvement is achieved by the CD-PAD framework without additional domain regularization reducing the BPCER at 1% APCER by 7.94% and BPCER at 5% APCER by 13.72%.

4.6 Discussion and Analysis

This section covers the analysis and ablation studies that went into developing the CD-PAD framework. First, we analyze the trade-off on performance on the source (visible) data when training at different durations on target imagery. Second, we consider the CD-PAD method with and without various configurations of the DDA subnetwork placed at varying depths in the base network architecture. Third, we evaluate the performance of the network when changing the dimensionality of the final embedding representation produced by the CD-PAD network.

4.6.1 Source-Target Trade Off Analysis

The key to CD-PAD is to find an IR-like embedding space that enhances the discriminability of visible imagery. In Section 4.5, it is shown that the single mode IR baselines significantly outperform the single mode visible baseline. We consider the NIR and thermal baselines as an upper limit of what can be achieved through cross

domain training. In Section 1.2 we used SSIM to show the domain gap between visible and infrared imaging modalities.

Table 4.5: Varying training epochs used in Target learning stage of CD-PAD for WMCA with NIR target domain without using the DDA subnet and finetuning LCNN layers

Target Epochs	BPCER @ 1% APCER	BPCER @ 5% APCER	AUC
1	68.52 ± 1.95	45.85 ± 0.94	0.884 ± 0.006
5	65.65 ± 1.21	52.27 ± 2.48	0.848 ± 0.003
10	55.36 ± 5.38	**35.73 ± 10.99**	**0.909 ± 0.032**
15	**52.49** ± 4.45	40.92 ± 1.69	0.908 ± 0.009
20	57.30 ± 2.76	38.41 ± 1.88	0.908 ± 0.016
25	64.17 ± 2.59	45.26 ± 10.40	0.884 ± 0.002

Table 4.6: Varying training epochs used in Target learning stage of CD-PAD for WMCA with Thermal target domain without the DDA subnet

Epochs Stage 1	BPCER @ 1% APCER	BPCER @ 5% APCER	AUC
1	86.66 ± 3.26	73.65 ± 3.85	0.763 ± 0.026
5	56.82 ± 13.37	41.46 ± 12.05	0.895 ± 0.033
10	73.83 ± 3.34	54.29 ± 1.73	0.881 ± 0.009
15	**46.23 ± 3.01**	**32.87 ± 6.23**	**0.917 ± 0.013**
20	54.16 ± 7.46	39.94 ± 6.72	0.905 ± 0.021
25	60.87 ± 4.43	39.08 ± 8.46	0.910 ± 0.012

In Table 4.5 and 4.6 results are shown for the CD-PAD method before adding the DDA subnet, instead layers of the Light CNN are made trainable to transform source representations. In both cases, performance declines beyond 15 epochs of initial classifier training on target data.

In Tables 4.8 and 4.7, the DDA subnet is incorporated into the CD-PAD framework to generate the results shown. The inclusion of the DDA subnetwork makes the overall approach less sensitive to the issue of overtraining on target imagery. Per-

Table 4.7: Varying training epochs used in Target learning stage of CD-PAD for WMCA with Thermal target domain with the DDA subnet in use

Target Epochs	BPCER @ 1% APCER	BPCER @ 5% APCER	AUC
1	40.10 ± 5.81	16.61 ± 5.31	0.945 ± 0.010
5	26.43 ± 4.02	11.55 ± 3.71	0.964 ± 0.010
10	**26.02 ± 0.28**	**9.72 ± 3.99**	**0.971 ± 0.009**
15	26.61 ± 8.64	14.62 ± 4.70	0.954 ± 0.013
20	27.48 ± 1.08	10.46 ± 2.39	0.969 ± 0.001
25	26.46 ± 0.47	12.87 ± 1.36	0.962 ± 0.001

Table 4.8: Varying training epochs used in Target learning stage of CD-PAD for WMCA with NIR target domain when using the DDA subnet

Target Epochs	BPCER @ 1% APCER	BPCER @ 5% APCER	AUC
1	24.04 ± 6.86	13.53 ± 3.83	0.960± 0.018
5	20.04 ± 2.30	12.70 ± 3.44	0.971 ± 0.008
10	**18.09 ± 5.50**	**12.57 ± 4.66**	**0.964 ± 0.015**
15	19.71 ± 1.07	11.64 ± 1.92	0.967 ± 0.001
20	28.94 ± 1.78	15.14 ± 1.37	0.955 ± 0.001
25	23.01 ± 2.90	8.87 ± 0.05	0.973 ± 0.004

formance peaks after 10 epochs, however continued training on target imagery only causes the results to plateau instead of the steady increase in error rate seen when trained without the subnet.

4.6.2 Subnetwork Ablation Study

A subnetwork ablation study was conducted to determine the optimal layer depth at which to insert a domain adaptive subnetwork into the visible channel of the CD-PAD network, and to evaluate different potential subnetwork architectures. The Light CNN network contains four max pooling layers that conclude each convolutional block. For each test, a trainable subnetwork is placed directly after one of the max pooling layers

to learn the transformation from the source to target domain. When a subnetwork is used for domain adaptation, all of the pre-trained layers in the network remain fixed during training.

Table 4.9 shows the effects of subnetwork type (residual or dense) and location when using the CD-PAD method. For this ablation study, additional regularizing loss functions are not implemented in the domain adaptive phase of training in order to highlight the change in performance that can be attributed to the network architecture alone.

For both architectures, the domain adaptive subnetwork shows the greatest effect when placed after the third pooling layer in the Light CNN. The most drastic improvements are seen in the lowest false positive rates where the CD-PAD network struggles without having the support of additional domain regularization. All of the final results presented for the CD-PAD framework are generated using the domain adaptive block at the third max pooling layer, which we refer to as the DDA subnetwork.

Table 4.9: Subnetwork ablation study

Network Details		Visible / Thermal		Visible / NIR	
Subnet Type	Layer	BPCER @1% APCER	BPCER @5% APCER	BPCER @1% APCER	BPCER @5% APCER
None	No DDA	41.69 ± 17.32	34.67 ± 15.99	64.32 ± 2.79	50.92 ± 4.71
Dense	Pool2	66.75 ± 8.45	53.32 ± 11.79	62.68 ± 9.99	71.47 ± 23.55
	Pool3	**29.64** ±17.89	**13.78** ± 7.11	**18.7** ± 1.77	**9.95** ± 1.31
	Pool4	66.61 ± 12.17	51.97 ± 19.99	71.47 ± 23.55	55.66 ± 19.44
Residual	Pool2	68.45 ± 10.99	45.11 ± 14.49	68.24 ± 12.9	50.11 ± 14.49
	Pool3	39.75 ± 14.09	18.99± 10.17	20.69 ± 3.22	10.37 ± 1.06
	Pool4	86.92 ± 10.77	66.21± 10.11	95.9 ± 1.05	79.64 ± 5.28

4.6.3 Embedding Dimensionality Study

In deep learning, high dimensionality is a topic of concern with respect to both model weights and data representation. The final layers of a classification model are gener-

ally comprised of one or more fully connected layers. Therefore, a high-dimensional feature embedding can result in over parameterization in the classifier. We studied the effect of increasing and decreasing the embedding dimensionality of the CD-PAD target and inference networks. The results of this study for visible to thermal CD-PAD are presented in Table 4.10, and the results for visible to NIR training are in Table 4.11.

Since changing the embedding size requires randomly initializing the final fully connected layer of the Light CNN base network, all of the embedding sizes considered, 128 through 1024, have randomly initialized weights generated using the He method (He et al. (2015)). We know how the CD-PAD network performs when the base network is fully pretrained, this provides additional insight into whether or not adapting additional parameters can help or harm performance.

Table 4.10: Varying the image embedding dimensionality for the thermal domain of WMCA

Embed. Size	Train Params	BPCER @ 1% APCER	BPCER @ 5% APCER	AUC
128	4,404,053	54.90 ± 27.18	**18.33 ± 7.39**	**0.926 ± 0.015**
256	8,600,149	**50.46 ± 23.49**	21.06 ± 8.65	0.884 ± 0.050
512	16,992,341	81.01 ± 5.34	38.90 ± 7.15	0.896 ± 0.020
1024	33,776,725	62.37 ± 14.44	26.04 ± 6.11	0.912 ± 0.029

For a point of comparison, the CD-PAD network with the DDA subnetwork that uses pre-trained weights in the fully connected layer only has 210,517 parameters that can be adapted over the course of both stages of training. It is clear that increasing the feature size to 1024 raises the over parameterization problem since BPCER performance at 1% APCER increases by 20.68% and 25.83% for thermal and NIR cross-domain training respectively compared against the “No DDA” baseline

Table 4.11: Varying the image embedding dimensionality for the NIR domain of WMCA

Embed. Size	Train Params	BPCER @ 1% APCER	BPCER @ 5% APCER	AUC
128	4,404,053	**61.64 ± 1.58**	11.97 ± 0.77	0.945 ± 0.007
256	8,600,149	70.09 ± 10.09	**10.49 ± 3.20**	**0.967 ± 0.011**
512	16,992,341	68.75 ± 3.95	11.01 ± 1.61	0.956 ± 0.005
1024	33,776,725	90.15 ± 4.37	50.75 ± 21.86	0.917 ± 0.026

from Table 4.9.

Chapter 5

CASIA-SURF Development

In this chapter, the additional analysis involved in applying the CD-PAD method to the CASIA-Surf dataset is expanded upon. Limitations of the data in the target domain necessitated additional experimentation with the CD-PAD method.

5.1 Single Mode Baselines

First, the baseline expectation for PAD performance must be established for each of the imaging modalities individually. Previously we have seen that infrared images are more discriminable for the PAD task using thermal and NIR from WMCA and NIR from MSSPOOF. Casia-Surf has significant quality differences, particularly in the NIR domain, so it is unknown if it is suitable for the CD-PAD approach.

Comparing the results from the single mode baselines, the NIR image domain does not show enough improvement over visible to provide a desirable "target" for cross-domain training. In Figure 5.1 the ROC curves show that for this dataset the NIR baseline has a higher BPCER for low values of APCER than the visible baseline. Adding a DDA subnetwork to the target stream improved the single mode PAD performance with NIR to a level that could make the NIR image representation a suitable target.

In Table 5.1 side by side comparisons are shown for the different single mode baselines considered for CASIA-Surf. Introducing the DDA subnetwork into the single mode networks provides a performance boost for NIR, but does not provide the same effect for visible imagery. CD-PAD*, the modified CD-PAD network with an additional "target" domain DDA subnetwork, is able to utilize the improvement in the NIR domain.

While CD-PAD* does slightly improve results with CASIA-Surf, this study indicates that for optimal cross-domain performance the quality of the "target" domain data should at least be similar to the "source".

Table 5.1: Results for single modal baselines on the Casia-Surf dataset

Method	BPCER @ 1% APCER	BPCER @ 5% APCER	AUC
MCCNN(NIR) George et al. (2020)	85.98± 0.07	65.93 ± 3.78	0.843 ± 0.028
MCCNN(Visible) George et al. (2020)	65.64 ± 12.79	39.35 ± 15.89	0.911 ± 0.048
NIR (with DDA)	19.42 ± 9.93	3.64 ± 2.41	0.991 ± 0.004
Visible (with DDA)	63.16 ± 11.01	35.66 ± 13.95	0.923 ± 0.042
CD-PAD	61.02 ± 14.37	26.18 ± 5.23	0.939 ± 0.015
CD-PAD*	**55.22 ± 4.20**	**21.94 ± 2.40**	**0.957 ± 0.003**

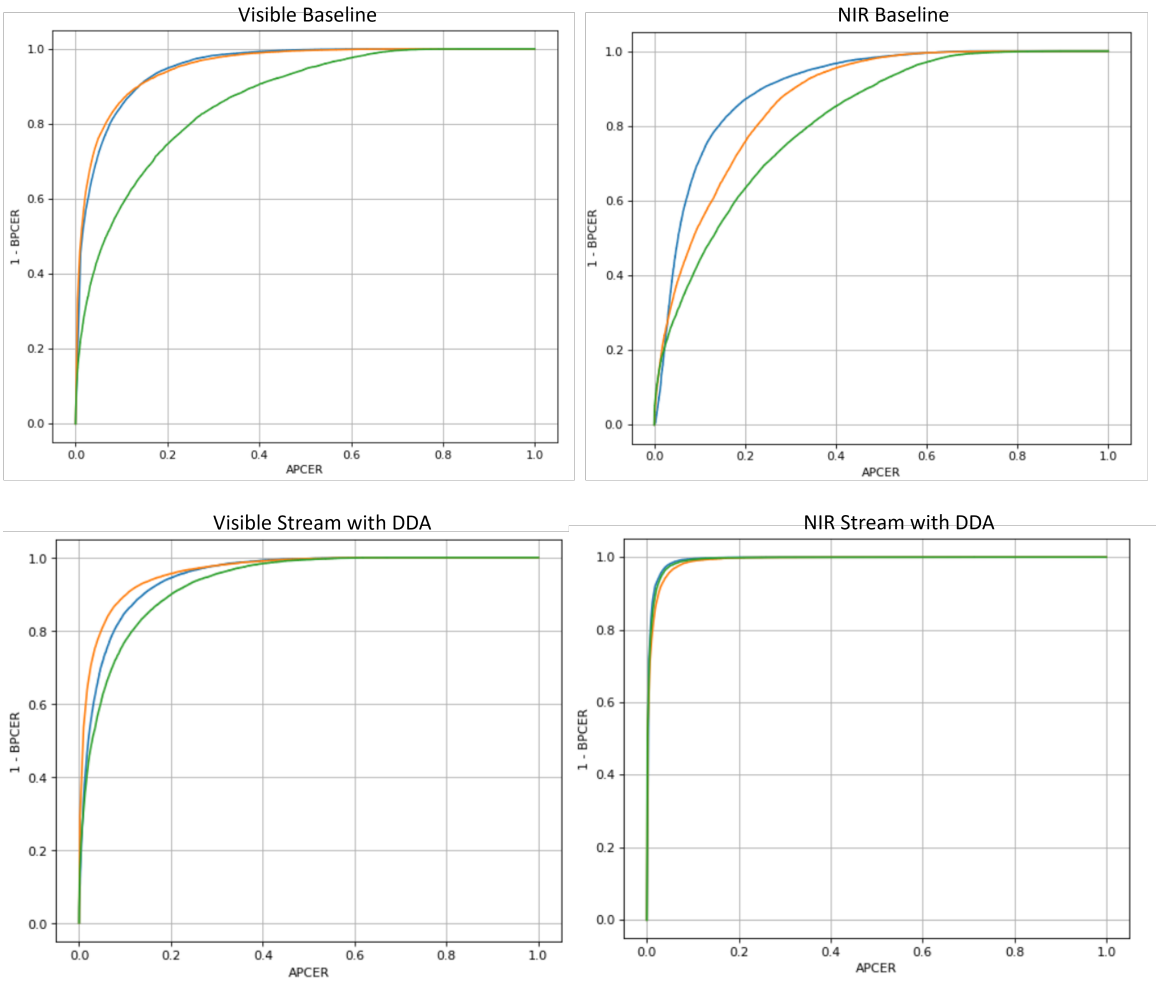


Figure 5.1: ROC curves for the different single mode configurations for CAIA-Surf. Adding an additional DDA subnetwork to the target stream

Chapter 6

Discussion

6.1 Data Challenges

As generalizability is a concern for any deep learning model, it is important to evaluate on multiple data sets to ensure that a new approach is not overly specific in its effectiveness. For this research we specifically sought out PAD datasets that contain mutliple imaging domains, namely visible and infrared.

6.1.1 CASIA-SURF

A primary concern with CASIA-Surf has to do with the relative scale of the source and target data. In Table 6.1 the average number of pixels per raw image is compared for the different subsets and split by domain. In biometric applications that use the face the typical scale comparison metric is distance between the eyes, however face landmark detection is primarily trained on visible imagery and this analysis could not be performed with the NIR data. On average an image from the source domain contains between three and four times as many pixels as an image in the target domain.

Examples of a random selection of input images for each domain are shown in Figure 6.1 where it is evident the lack of fine detail contained in some of the target

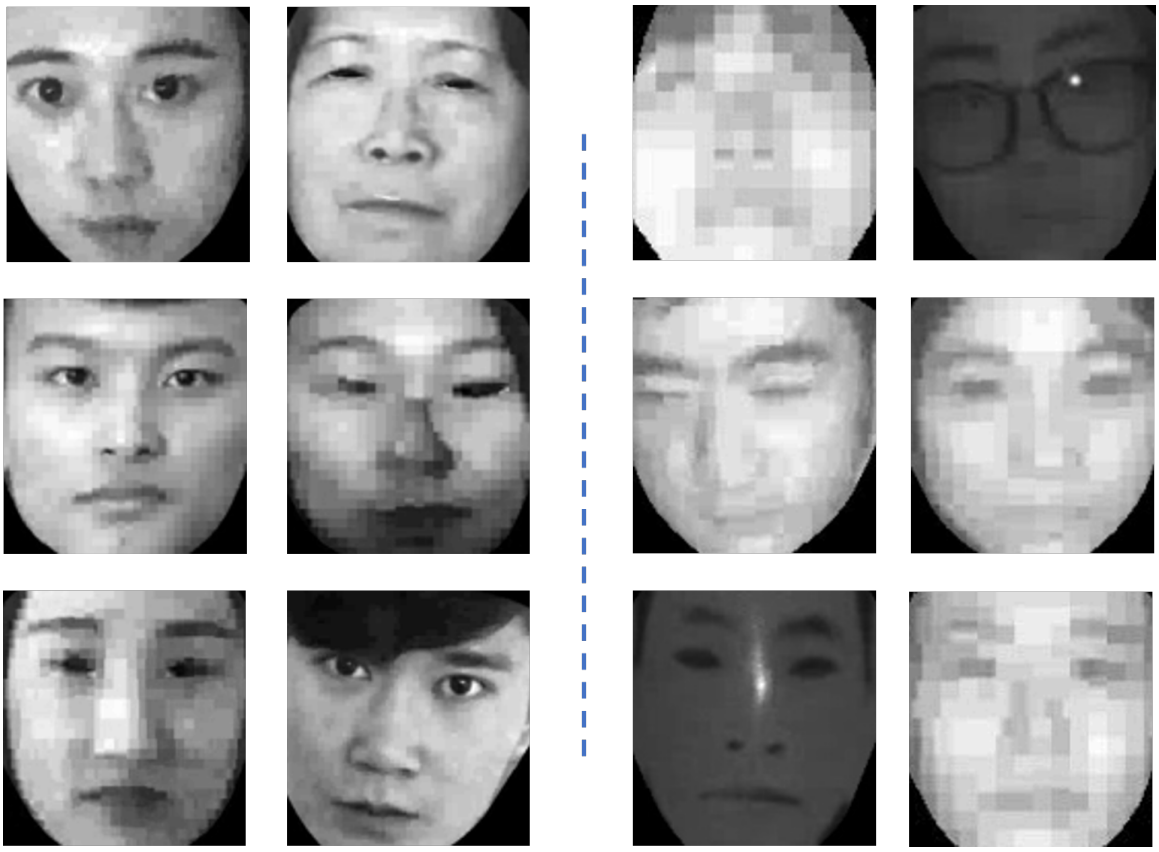


Figure 6.1: Raw images in the CASIA-Surf dataset vary in size within both spectral domains, however the scale issues are more pronounced in the NIR target domain. Left: Examples of pre-processed visible images. Right: Examples of pre-processed NIR images showing higher degree of pixelation.

images. This is less ideal for an application like CD-PAD where the NIR representation is used to enhance the Visible representation instead of acting as supplementary information at evaluation time.

Table 6.1: Results for single modal baselines on the Casia-Surf CeFA dataset

subset	Visible	NIR	Vis/NIR
Train	76415.35 ± 25258.45	20663.64 ± 6553.67	3.69
Test	78705.52 ± 26226.79	20641.35 ± 6672.19	3.81
Dev	83058.53 ± 28351.04	21656.62 ± 7127.36	3.83

Chapter 7

Conclusions

We proposed a new domain adaptation framework called CD-PAD that utilized mutli-modal data during training to improve visible based PAD for face recognition systems. we proposed a new CD-PAD framework, a domain adaptation approach to PAD for face recognition.

The goal of this framework was to utilize multi-modal face data during training to improve PAD when deployed on a facial recognition system that is only comprised of sensors and imagery in the visible spectrum.

qualitative analysis indicates an improvement in the clustering and separability of the bonafide and attack feature space.

To this end, we introduced (1) a new CD-PAD framework that increases the separability of bonafide and presentation attacks using only visible spectrum imagery, (2) an IDR technique for enhanced PAD and stability during optimization, and (3) a DDA subnetwork to transform representations between visible and infrared domains. We found that our CD-PAD framework was able to significantly reduce the BPCER @ 1% APCER by 57.46%, 62.17% and 29.37% on the WMCA (NIR), WMCA (thermal), and MSSpoof (NIR) protocols. Moreover, we found that our proposed IDR resulted in better PAD performance than previous MMD and DIL techniques. The results imply that the CD-PAD framework is capable of providing very discriminative PAD

while reducing the number/type of operation sensors, which enables less complex and more cost efficient PAD systems.

Additional experiments on the CASIA-Surf dataset shows that CD-PAD does require suitable image quality in the target domain. The modification of CD-PAD* still provides a modest improvement of the BPCER at 1% APCER by 7.94% and BPCER at 5% APCER by 13.72%.

Bibliography

Akshay Agarwal, Daksha Yadav, Naman Kohli, Richa Singh, Mayank Vatsa, and Afzel Noore. Face presentation attack with latex masks in multispectral videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

Akshay Agarwal, Akarsha Sehwag, Richa Singh, and Mayank Vatsa. Deceiving face presentation attack detection via image transforms. In *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*, pages 373–382, 2019. doi: 10.1109/BigMM.2019.00018.

Shervin Rahimzadeh Arashloo, Josef Kittler, and William Christmas. An anomaly detection approach to face spoofing detection: A new formulation and evaluation protocol. *IEEE Access*, 5:13868–13882, 2017. doi: 10.1109/ACCESS.2017.2729161.

Yousef Atoum, Yaojie Liu, Amin Jourabloo, and Xiaoming Liu. Face anti-spoofing using patch and depth-based cnns. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pages 319–328, 2017. doi: 10.1109/BTAS.2017.8272713.

Sushil Bhattacharjee, Amir Mohammadi, and Sébastien Marcel. Spoofing deep face recognition with custom silicone masks. In *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–7, 2018. doi: 10.1109/BTAS.2018.8698550.

Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a" siamese" time delay neural network. *Advances in neural information processing systems*, 6, 1993.

Ivana Chingovska, Nesli Erdogmus, André Anjos, and Sébastien Marcel. *Face Recognition Systems Under Spoofing Attacks*, pages 165–194. Springer International Publishing, 2016. ISBN 978-3-319-28501-6. doi: 10.1007/978-3-319-28501-6_8. URL `https://doi.org/10.1007/978-3-319-28501-6_8`.

Tiago de Freitas Pereira, André Anjos, and Sébastien Marcel. Heterogeneous face recognition using domain specific units. *IEEE Transactions on Information Forensics and Security*, 14(7):1803–1816, 2019. doi: 10.1109/TIFS.2018.2885284.

G. B. de Souza, D. F. da Silva Santos, R. G. Pires, A. N. Marana, and J. P. Papa. Deep texture features for robust face spoofing detection. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 64(12):1397–1401, 2017.

Tejas I. Dhamecha, Aastha Nigam, Richa Singh, and Mayank Vatsa. Disguise detection and face recognition in visible and thermal spectrums. In *2013 International Conference on Biometrics (ICB)*, pages 1–8, 2013. doi: 10.1109/ICB.2013.6613019.

N. Erdogmus and S. Marcel. Spoofing 2d face recognition systems with 3d masks. In *2013 International Conference of the BIOSIG Special Interest Group (BIOSIG)*, pages 1–8, 2013.

Cedric Nimpa Fondje, Shuowen Hu, Nathaniel J. Short, and Benjamin S. Riggan. Cross-domain identification for thermal-to-visible face recognition, 2020.

A. George and S. Marcel. Learning one class representations for face presentation attack detection using multi-channel convolutional neural networks. *IEEE*

Transactions on Information Forensics and Security, 16:361–375, 2021. doi: 10.1109/TIFS.2020.3013214.

A. George, Z. Mostaani, D. Geissenbuhler, O. Nikisins, A. Anjos, and S. Marcel. Biometric face presentation attack detection with multi-channel convolutional neural network. *IEEE Transactions on Information Forensics and Security*, 15:42–55, 2020.

Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex J. Smola. A kernel method for the two-sample-problem. In B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 513–520. MIT Press, 2007. URL `http://papers.nips.cc/paper/3110-a-kernel-method-for-the-two-sample-problem.pdf`.

Guodong Guo and Na Zhang. A survey on deep learning based face recognition. *Computer Vision and Image Understanding*, 189:102805, 2019. ISSN 1077-3142. doi: https://doi.org/10.1016/j.cviu.2019.102805. URL `http://www.sciencedirect.com/science/article/pii/S1077314219301183`.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Guillaume Heusch and Sébastien Marcel. Pulse-based features for face presentation attack detection. In *2018 IEEE 9th International Conference on Biometrics Theory,*

Applications and Systems (BTAS), pages 1–8, 2018. doi: 10.1109/BTAS.2018.8698579.

Guillaume Heusch, Anjith George, David Geissbühler, Zohreh Mostaani, and Sébastien Marcel. Deep models and shortwave infrared information to detect face presentation attacks. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2(4):399–409, 2020. doi: 10.1109/TBIOM.2020.3010312.

Shuowen Hu, Nathaniel Short, Benjamin S. Riggan, Matthew Chasse, and M. Saquib Sarfraz. Heterogeneous face recognition: Recent advances in infrared-to-visible matching. In *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, pages 883–890, 2017. doi: 10.1109/FG.2017.126.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

ISO/IEC 30107-3:2017. Information technology — Biometric presentation attack detection. Standard, International Organization for Standardization, Geneva, CH, 2017.

Fangling Jiang, Pengcheng Liu, and Xiangdong Zhou. Multilevel fusing paired visible light and near-infrared spectral images for face anti-spoofing. *Pattern Recognition Letters*, 128:30–37, 2019. ISSN 0167-8655. doi: https://doi.org/10.1016/j.patrec.2019.08.008. URL `https://www.sciencedirect.com/science/article/pii/S016786551830583X`.

Amin Jourabloo, Yaojie Liu, and Xiaoming Liu. Face de-spoofing: Anti-spoofing via noise modeling. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015.

Brendan Klare and Anil K. Jain. Heterogeneous face recognition: Matching nir to visible light images. In *2010 20th International Conference on Pattern Recognition*, pages 1513–1516, 2010. doi: 10.1109/ICPR.2010.374.

Brendan F. Klare and Anil K. Jain. Heterogeneous face recognition using kernel prototype similarities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(6):1410–1422, 2013. doi: 10.1109/TPAMI.2012.229.

Ketan Kotwal, Sushil Bhattacharjee, and Sébastien Marcel. Multispectral deep embeddings as a countermeasure to custom silicone mask presentation attacks. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 1(4):238–251, 2019. doi: 10.1109/TBIOM.2019.2939421.

Ketan Kotwal, Zohreh Mostaani, and Sébastien Marcel. Detection of age-induced makeup attacks on face recognition systems using multi-layer deep features. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2(1):15–25, 2020. doi: 10.1109/TBIOM.2019.2946175.

José Lezama, Qiang Qiu, and Guillermo Sapiro. Not afraid of the dark: Nir-vis face recognition via cross-spectral hallucination and low-rank embedding. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6807–6816, 2017. doi: 10.1109/CVPR.2017.720.

H. Li, P. He, S. Wang, A. Rocha, X. Jiang, and A. C. Kot. Learning generalized deep feature representation for face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 13(10):2639–2652, 2018. doi: 10.1109/TIFS.2018.2825949.

Ajian Liu, Zichang Tan, Jun Wan, Sergio Escalera, Guodong Guo, and Stan Z. Li. Casia-surf cefa: A benchmark for multi-modal cross-ethnicity face anti-spoofing. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1179–1187, January 2021.

Yaojie Liu, Amin Jourabloo, and Xiaoming Liu. Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Yaojie Liu, Joel Stehouwer, Amin Jourabloo, and Xiaoming Liu. Deep tree learning for zero-shot face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Mingsheng Long, Jianmin Wang, Guiguang Ding, Jiaguang Sun, and Philip S. Yu. Transfer feature learning with joint distribution adaptation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2013.

Amir Mohammadi, Sushil Bhattacharjee, and Sébastien Marcel. 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)*, pages 1001–1005, 2020. doi: 10.1109/ICASSP40776.2020.9053685.

Saeid Motiian, Marco Piccirilli, Donald A. Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

Jukka Määttä, Abdenour Hadid, and Matti Pietikäinen. Face spoofing detection from single images using micro-texture analysis. In *2011 International Joint Conference on Biometrics (IJCB)*, pages 1–7, 2011. doi: 10.1109/IJCB.2011.6117510.

Olegs Nikisins, Anjith George, and Sébastien Marcel. Domain adaptation in multi-channel autoencoder based features for robust face anti-spoofing. In *2019 International Conference on Biometrics (ICB)*, pages 1–8, 2019. doi: 10.1109/ICB45273.2019.8987247.

T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002. doi: 10.1109/TPAMI.2002.1017623.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL `http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf`.

Daniel Perez-Cabo, David Jimenez-Cabello, Artur Costa-Pazo, and Roberto J. Lopez-Sastre. Deep anomaly detection for generalized face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

Domenick Poster, Matthew Thielke, Robert Nguyen, Srinivasan Rajaraman, Xing Di, Cedric Nimpa Fondje, Vishal M. Patel, Nathaniel J. Short, Benjamin S. Riggan,

Nasser M. Nasrabadi, and Shuowen Hu. A large-scale, time-synchronized visible and thermal face dataset. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1559–1568, January 2021.

R. Raghavendra, Kiran B. Raja, Sushma Venkatesh, and Christoph Busch. Extended multispectral face presentation attack detection: An approach based on fusing information from individual spectral bands. In *2017 20th International Conference on Information Fusion (Fusion)*, pages 1–6, 2017. doi: 10.23919/ICIF.2017.8009749.

R. Raghavendra, Narayan Vetrekar, Kiran B. Raja, R. S. Gad, and Christoph Busch. Detecting disguise attacks on multi-spectral face recognition through spectral signatures. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 3371–3377, 2018. doi: 10.1109/ICPR.2018.8545076.

A. Rozantsev, M. Salzmann, and P. Fua. Beyond sharing weights for deep domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41 (4):801–814, 2019.

Rui Shao, Xiangyuan Lan, Jiawei Li, and Pong C. Yuen. Multi-adversarial discriminative deep domain generalization for face presentation attack detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Holger Steiner, Andreas Kolb, and Norbert Jung. Reliable face anti-spoofing using multispectral swir imaging. In *2016 International Conference on Biometrics (ICB)*, pages 1–8, 2016. doi: 10.1109/ICB.2016.7550052.

Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

Laurens van der Maaten and Geoffrey Hinton. Viualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 11 2008.

Guoqing Wang, Hu Han, Shiguang Shan, and Xilin Chen. 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 (CVPR)*, June 2020.

Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861.

Karl Weiss, Taghi M. Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big Data*, 3(1):9, May 2016. ISSN 2196-1115. doi: 10.1186/s40537-016-0043-6. URL `https://doi.org/10.1186/s40537-016-0043-6`.

X. Wu, R. He, Z. Sun, and T. Tan. A light cnn for deep face representation with noisy labels. *IEEE Transactions on Information Forensics and Security*, 13(11): 2884–2896, 2018.

Fei Xiong and Wael AbdAlmageed. Unknown presentation attack detection with face rgb images. In *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–9, 2018. doi: 10.1109/BTAS.2018.8698574.

S. Zhang, A. Liu, J. Wan, Y. Liang, G. Guo, S. Escalera, H. J. Escalante, and S. Z. Li. Casia-surf: A large-scale multi-modal benchmark for face anti-spoofing. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2(2):182–193, 2020a. doi: 10.1109/TBIOM.2020.2973001.

Y. Zhang, M. Zhao, L. Yan, T. Gao, and J. Chen. Cnn-based anomaly detection for face presentation attack detection with multi-channel images. In *2020 IEEE International Conference on Visual Communications and Image Processing (VCIP)*, pages 189–192, 2020b. doi: 10.1109/VCIP49819.2020.9301818.