Multispectral Deep Embeddings As a Countermeasure To Custom Silicone Mask Presentation Attacks

September 2, 2019

Abstract

This work focuses on detecting presentation attacks (PA) mounted using custom silicone masks. Face recognition (FR) systems have been shown to be highly vulnerable to PAs based on such masks [1, 2]. Here we explore the use of multispectral data (color imagery, near infrared (NIR) imagery and thermal imagery) for face presentation attack detection (PAD), specifically against the custom silicone mask attacks. Using a new dataset (XCSMAD) representing 21 custom made masks, we establish the baseline performance of several commonly used face-PAD methods, on the different imaging channels. Considering thermal imagery in particular, our experiments show that low-cost thermal imaging devices are as effective in face-PAD as more expensive thermal cameras, for mask-based attacks. This result reinforces the case for the use of thermal data in face-PAD.

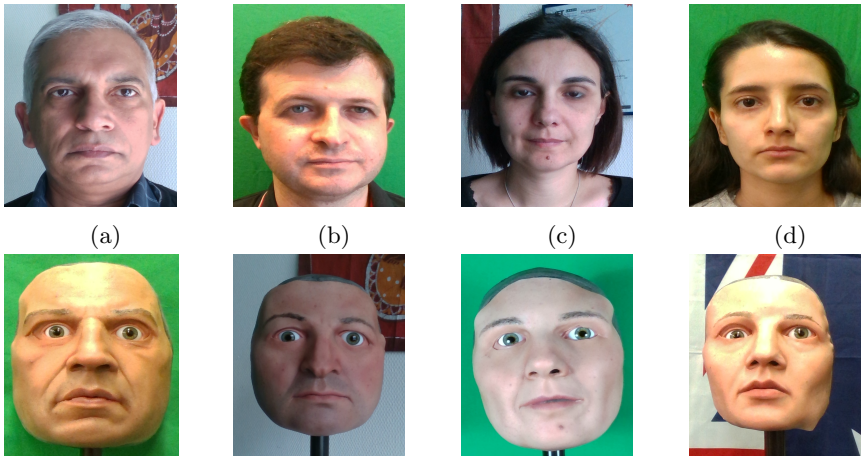
We also demonstrate that fusing information from multiple channels leads to significant improvement in face-PAD performance. Finally, we propose a new approach to face-PAD of custom silicone masks using a convolutional neural network (CNN). On individual spectral channels, the proposed approach achieves state-of-the-art results. Using multispectral-fusion, the proposed CNN-based method significantly outperforms the baseline methods. The new dataset and source-code for our experiments is freely available for research purposes.

Keywords: Face Presentation Attack Detection (PAD), Biometrics, Custom Silicone Masks, Multispectral face-PAD, Convolutional Neural Network (CNN), Deep Learning, CNN Embeddings, Feature Fusion, Score Fusion, XCSMAD

1 Introduction

Presentation attacks (PA), are the most common form of attacks on an FR system. Various studies [3, 4] have shown that state-of-the-art face recognition (FR) systems, while achieving near perfect FR performance even in challenging

scenarios, are highly vulnerable to PAs. Therefore, to have a trustworthy face-based identity-verification system, it is imperative to pair a FR method with an appropriate presentation attack detection (PAD) method.



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 1: Examples of custom silicone mask-based PAs. (a)–(d) *bona fide* samples; (e)–(h) custom-mask attack presentations. The presentations have been captured against different backgrounds under controlled illumination.

There are two kinds of face-PAs: (1) *impersonation* and (2) *obfuscation*. In an impersonation attack, the identity of a subject (enrolled in a FR system) is attacked using a PA. An obfuscation attack happens when a user attempts to deceive the FR system to *avoid* being recognized. In this work we focus on impersonation PAs, which fall into three broad categories (depending on the medium used to mount the PA): (1) print attacks, (2) digital replay attacks, and (3) 3D mask attacks [5]. In attacks of the first two categories, the instrument used to perform the attack is inherently two-dimensional (2D). Therefore, such attacks are collectively called 2D attacks. PAs in the third category are created using custom masks.

Most face-PAD research so far has been focused on detecting 2D impersonation PAs [6]. Very few studies so far have considered the challenge of detecting 3D-mask PAs. One reason why, is the difficulty of creating realistic custom masks. Until recently, realistic masks were artisanal products, time-consuming, and therefore expensive to produce. Recent advances in depth sensing technology, and the rise of 3D printers have facilitated the production of rigid custom masks at a reasonable cost ($<$ USD500). Past face-PAD studies related to mask-based impersonation attacks [7, 8] have considered primarily custom rigid masks.

In this paper we present the first data-driven analysis of countermeasures to impersonation PAs constructed using *custom silicone* masks. Compared to rigid masks, these flexible masks have a much more realistic appearance (Fig. 1).

Previous studies have shown that FR systems are highly vulnerable to custom-silicone-mask based PAs [1, 2]. One underlying reason for this vulnerability is that these FR systems are designed to operate on visible-light (RGB-color) imagery. The use of imagery in various near infrared (NIR) wavelength bands is particularly effective in detecting various kinds of PAs, including rigid mask based PAs [3]. In this study, we also explore the use of imagery in other wavelength bands, specifically in NIR as well as long-wave infrared (LWIR, *i.e.*, thermal) bands in detecting custom silicone mask based impersonation PAs. Good thermal cameras are typically quite expensive ($\approx$ USD 10,000). Recently low-cost ($\approx$ USD 500) thermal cameras have become available. Compared to the expensive thermal cameras, the low-cost thermal cameras have relatively lower image resolution, and lower thermal accuracy and sensitivity, as they lack adequate sensor-cooling subsystems. In this study, we also investigate whether low-cost thermal cameras can reliably be used for detecting custom flexible mask based PAs.

The present study is based on a new dataset of *bona fide* presentations and custom silicone mask based PAs. The PAs in this dataset come from 21 masks corresponding to 17 subjects. (For most subjects only one custom mask has been created, but multiple custom silicone masks are available for some subjects.) Each presentation has been captured in four imaging channels: color, NIR, high quality LWIR, and low quality LWIR. The main contributions of this paper are:

1. A new PAD method for custom silicone mask detection based on a CNN. Experimental results demonstrate the efficacy of the proposed method over baseline methods.
2. Using extended-range (ER) imagery (color, NIR, and LWIR), we demonstrate that fusing information from the different channels leads to a better face-PAD performance than single-channel approaches.
3. Performance analysis of several well known baseline face-PAD methods in detecting custom silicone masks. We test these methods on color, NIR, and LWIR images.
4. A new multispectral dataset, named *XCSMAD*, for face-PAD studies involving custom silicone mask attacks[1].
5. We also show that data from low-cost thermal cameras can produce face-PAD performance comparable to data from more expensive thermal cameras. This result reinforces the case for using thermal images for face-PAD.
6. Vulnerability-analysis of LightCNN based FR method ([9]) to the custom-mask based attacks in the new dataset.

The discussion about previous research, presented in Section 2, develops the context for the present work. Details of mask construction, data acquisition and protocols for PAD experiments are presented in Section 3. Section 4 discusses the proposed CNN-based face-PAD method. We describe the experiments and discuss results in Section 5. A summary of the work, and conclusions drawn from it, are presented in Section 6.

[1]Dataset: `https://www.idiap.ch/dataset/xcsmad/`

2 Related Work

This work has three major aspects: custom-made 3D mask PAs, ER imagery for PAD, and PAD using deep networks. In this section we review of the relevant literature in these three areas. These related works are also summarized in Table 1.

2.1 3D-Mask PAD

Two broad categories of 3D-masks have been considered in the PAD literature: rigid masks and flexible masks. Before delving into countermeasures for custom mask based PAs, the threat posed by such PAs should be quantified. Erdogmus and Marcel [7] presented the first study demonstrating the vulnerability of FR systems to PAs made using custom 3D masks. Their study is based on the 3DMAD dataset [7], that has since been widely used to benchmark various 3D-mask PAD methods.

Bhattacharjee *et al.* [1] have recently shown that several state-of-the-art CNN-based FR systems are highly vulnerable to custom silicone mask based impersonation PAs. Previous face-PAD studies involving flexible masks [10, 11] have used generic, not custom-made masks, to address the challenge of obfuscation PAs. Therefore, these works cannot be directly compared with the present study.

PAD methods designed for 2D attacks have also been applied to detect 3D-mask attacks. Erdogmus and colleagues [7, 12] have used local binary pattern (LBP) histograms [13] to detect mask-based PAs. Liu *et al.* [8] have also explored the use of LBP features for detecting PAs in the 3DMAD. On the same dataset, Lina and Ramavel [14] have achieved lower error rates than [7] by using binary statistical image features (BSIF). 'Haralick features', texture descriptors derived from co-occurrence matrices, have also been used to detect PAs in the 3DMAD [15]. In the present work, for the first time we will test several such countermeasures, developed for 2D PAs, on custom silicone-mask based PAs.

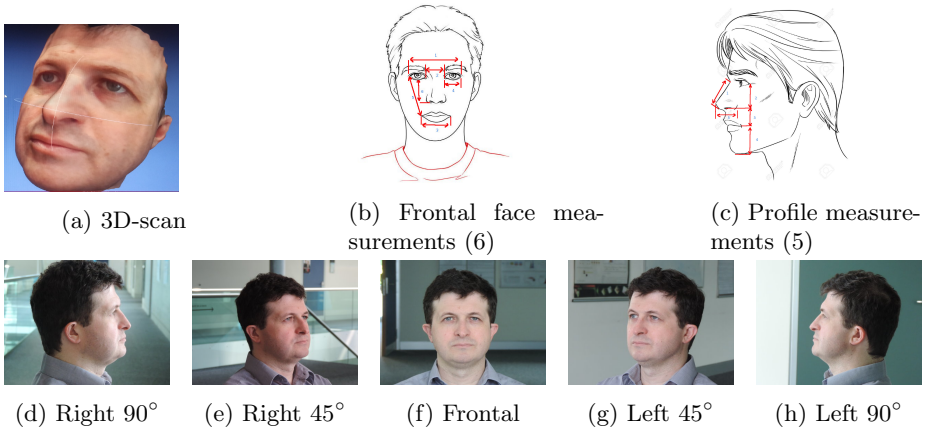
One class of 3D-mask PAD algorithms consists of approaches based on *remote photoplethysmography* (rPPG): a technique for optically tracking the cardiac-synchronous changes in blood-flow measured, say, over the facial region of a subject. The first work to successfully use rPPG for face-PAD was published by Li *et al.* [16]. In [17], rPPG signals from several disjoint facial zones have been combined to detect mask based PAs. Nowara *et al.* [18] have also used rPPG signals from different regions including the background, to determine whether a presentation is *bona fide*. As Liu *et al.* [17] point out, that although combining rPPG signals from different regions may suppress certain kinds of noise in the signal, such an approach can also reinforce other kinds of noise, notably noise due to camera-motion. Heusch and Marcel [19] have shown that long-term spectral statistics (LTSS) features [20] extracted from rPPG signals outperform other rPPG-based face-PAD methods [16, 18] on several datasets. The face-PAD accuracy achieved in their experiments, however, range from 13% to 25%. These results are far from the state-of-the-art accuracies achieved on

the respective datasets using other (non-rPPG) approaches.

The attraction of rPPG based face-PAD lies in the fact that only the *bona fide* class is modelled. Therefore, such a method can be effective against previously unseen classes of PAs. The main drawback, however, is that accurate rPPG signals are notoriously difficult to extract, unless the presentation-videos are collected under strictly controlled conditions of illumination and subject-motion.

Table 1: Summary of recent research activities.

Research Objective	**Ref.**	**Key Techniques/Contributions**	**Dataset(s)**
Vulnerability Analysis	[7]	Vuln. of 3-D custom masks	3DMAD
	[3]	Vuln. analysis in ER image-domain	EMSPAD (7 MS bands)
	[1]	Vuln. of CNN-based FR systems to 3D custom silicone masks	CSMAD
3D Mask Face-PAD	[7]	local binary patterns (LBP)	3DMAD
	[12]	LBP	Proprietary
	[8]	LBP	HKBU-MARs
	[14]	BSIF (binary statistical image features)	3DMAD
	[15]	Haralick features	3DMAD
ER imaging Face-PAD	[21]	Image fusion, Score fusion	
	[10]	Haralick features, BSIF, LBP, Local Phase Quantization (LPQ), Histogram of Gradients (HoG)	MLFP
Deep Learning-based Face-PAD	[22]	Transfer Learning	REPLAY-ATTACK, CASIA
	[23]	Fusion (CNN + eye-blink)	REPLAY-ATTACK, CASIA
	[24]	Fusion (CNN + multiscale LBP)	CASIA, NUAA
	[25]	Autoencoder, Shearlets, optical flow	REPLAY-ATTACK, CASIA
	[26]	LiveNet	REPLAY-ATTACK, CASIA
	[11]	Deep dictionary	SMAD
	[27]	Maximum Mean Discrepancy loss function	REPLAY-ATTACK, CASIA, MSU
	[28]	LSTM	REPLAY-ATTACK, CASIA, 3DMAD



(a) 3D-scan (b) Frontal face measurements (6) (c) Profile measurements (5)

(d) Right 90° (e) Right 45° (f) Frontal (g) Left 45° (h) Left 90°

Figure 2: Examples of facial data collected for mask construction. (a) Rendering of 3D face scan. The red double-headed arrows in (b) and (c) indicate the facial measurements collected for the purpose of mask construction (six frontal and five profile measurements are recorded). For each mask-subject (target), the manufacturer is also provided with photographs of various facial poses (d)–(h).

2.2 Extended-Range Imagery for Face-PAD

PAs are continually improving in quality (resolution, color fidelity, and so on, for 2D PAs, as well as realism of masks), and the distinction between the two classes of presentations in VIS imagery is becoming increasingly subtle. Several recent works have proposed face-PAD methods relying on *extended range* (ER) imagery, that is imagery in wavelengths covering the range from visible light (VIS) to LWIR. Ramachandra *et al.* [3] have analyzed the vulnerability of FR systems in seven wavelength bands covering the VIS and NIR portion of the spectrum where they consider photo-print based attacks. Subsequently [21], they have also proposed two PAD approaches based on the seven-band multi-spectral images: image-fusion and score-fusion. Their experiments show that the score-fusion approach performs better than the image-fusion approach.

Kanzawa *et al.* [29], working on driver assistance systems for vehicles, showed that human skin could be reliably detected, even at significant distances, by combining images captured in three NIR bands– 870nm, 970nm, and 1050nm– with VIS images. Their key finding was that the reflectance of human skin dips sharply around 970nm. In the context of biometrics two separate works, Bourlai [30] and Steiner *et al.* [31], have proposed multispectral short-wave infrared (SWIR) cameras to distinguish human skin from other materials.

Existing studies using ER-imagery for face-PAD have mainly focused on 2D PAs. The masks present a fundamentally different challenge to FR systems than 2D PAs. Bhattacharjee and Marcel [32] have shown that whereas 2D PAs can generally be easily detected with NIR imagery using relatively low-cost devices, such approaches cannot easily detect custom-mask based PAs. The reason is that the zeroth order and first-order statistics of mask images in both VIS and

NIR imaging domains are quite similar to those of *bona fide* presentations. In thermal images, *bona fide* presentations produce bright facial images, whereas masks, being significantly colder than the average body temperature, result in very dark (low intensity) face images. In other words, in the thermal domain the two kinds of presentations show significantly different low order statistics. Therefore, thermal imagery can be used to reliably detect 3D-mask PAs [32]. We note here that whereas cameras for VIS and NIR imagery are eminently affordable, SWIR and LWIR (thermal) cameras are usually very expensive. Relatively low-cost thermal cameras, such as from Flir One (`www.flir.com/flirone`) and Seek Thermal (`www.thermal.com`) have become available in past decade. SWIR cameras, however, still remain unaffordable for most applications.

2.3 Face-PAD using Deep Networks

Li *et al.* [22] first applied transfer-learning to adapt the VGG-Face [33] network, a benchmark CNN for FR, for 2D PAD. Their results on Replay-Attack dataset are comparable to those obtained using IQM features [34], but at a much higher computational cost. Patel *et al.* [23] fuse the decisions of two classifiers- one based on frame-based texture features extracted using a CNN, and the other relying on eye-blink detection, to detect PAs. Their method shows improved cross-dataset classification results compared to their chosen baseline methods. More recently, Nguyen *et al.* [24] have also proposed a similar approach, combining handcrafted (multiscale LBP (MLBP)) features with CNN embeddings for face-PAD. They show that combining the two kinds of features leads to better PAD performance than using CNN-embeddings only. Whereas Patel *et al.* [23] take a decision-fusion approach, Nguyen *et al.* have reported results using feature-fusion and score-fusion to combine the two channels of information. It is not pertinent to compare the two works, as Patel *et al.* [23] focus on cross-dataset performance whereas Nguyen *et al.* explore the idea of supplementing CNN-derived features with handcrafted features.

Rehman *et al.* [26] hypothesize that CNN based face-PAD methods show poor generalization in cross-dataset scenarios because of the over-learning induced by the use of small training sets. They have trained a CNN (VGG-11) using continuous randomization, a data-augmentation method, to train the network. Their experiments show that their CNN ('LiveNet') achieves lower error rates than previous works in cross-dataset tests.

In search of improved cross-camera generalization, Li *et al.* [27] have recently proposed the use of Maximum Mean Discrepancy (MMD) [35] as the loss function for training a 3D-CNN, which extracts both spatial and temporal characteristics of input videos. They also demonstrate that use of domain generalization [36] can improve the cross-camera generalization performance of several previously proposed face-PAD methods.

The works discussed above have applied CNN based methods for detecting 2D PAs. Very few works have included custom-made 3D-mask based attacks in their studies. Sun *et al.* [28] have investigated several deep network architectures, including stacked CNNs as well as CNN+LSTM networks, where the

output of the CNN is processed by a long short-term memory (LSTM) network, to model the temporal characteristics of the input. Their experiments show that most of these architectures can achieve near-perfect PAD performance on the 3DMAD dataset.

Shao *et al.* [37] have proposed a method for detecting rigid-mask based PAs. They use a pre-trained VGG network[2] to extract several channels of texture-features from different regions of the face. The texture-features are analyzed, channel-wise, over the sequence of video-frames, to learn the dynamic modulation of the texture-information in each channel. This analysis is used to characterize micro-motion of facial regions, which is used as a cue to discriminate between *bona fide* and 3D-mask presentations. This method outperforms the chosen baseline methods over two datasets: 3DMAD, and the SUP dataset [8].

3 Data Description

For this study we have used a new dataset named *XCSMAD* (eXtended Custom Silicone Mask Attack Dataset). Participants of the study have played three different roles: (1) *target*: person for whom custom-masks have been created; (2) *attacker*: subject who attacks a target's identity by wearing a custom-mask of the target, and (3) bona fide *subject*: person who makes a *bona fide* presentation. The dataset consists of *bona fide* presentations corresponding to 72 *bona fide* subjects, and attack presentations using 21 masks. The data collection process for our experiments is presented here. We start by describing the custom-masks used in this study, and the devices used for recording the *bona fide* and attack presentations that comprise the XCSMAD.

3.1 Custom-Mask Manufacturing Process

A total of 21 custom-masks have been used in this study. These masks have been manufactured by Nimba Creations Ltd., a special-effects company, at a cost of approximately USD 4000 per mask. For each target, the manufacturer was provided with the following data: (1) 3D-scan of the face collected using a Realsense SR300 camera, (2) physical measurements of facial features, and (3) color facial photographs from different points of view (frontal, lateral, and diagonal views). Figure 2 illustrates the various pieces of data provided to the manufacturer for each custom-mask.

The manufacturer starts by creating a 3D cast of the target's head (see Fig. 3). Such a cast is then used to create the silicone mask. The raw mask, of the kind shown in Fig. 4(a), undergoes a laborious manual finishing process where the appropriate skin color as well as other facial texture features (*e.g.*, beard, eyebrows, facial make-up and so on) are manually applied to the mask. Figure 4(c) shows an example of a finished mask.

For each mask, the manufacturer has also provided a bespoke matching support. Each support is shaped to roughly match the face-shape of the target,

[2] www.robots.ox.ac.uk/~vgg/research/very_deep/

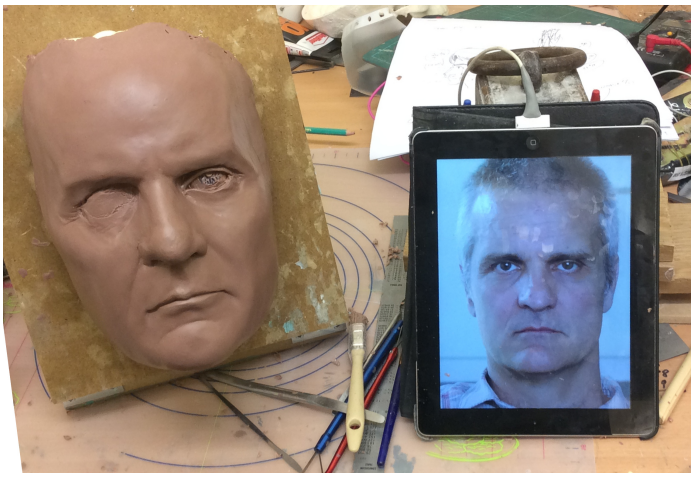


Figure 3: Example of cast made during the custom-mask manufacturing process.

(a) (b) (c)

Figure 4: Different stages in the process of creating a custom silicone mask. (a) Raw silicone mask; (b) intermediate stage; and (c) finished mask.

and contains holes for eye-sockets. In addition the manufacturer has supplied us with several sets of synthetic eye-balls with different iris colors, as well as silicone eye-sockets for mounting the eye-balls in the bespoke supports. We refer to these masks as *half-masks*, because each mask corresponds to only the facial region of the subject, and not the entire head. The inner surface of each mask is coated with a layer of glue which helps to hold the mask in position when worn. The masks are manufactured with holes cut-out for the eyes.

3.2 Data Acquisition

The data used in this study has been captured using three cameras, namely (1) Intel RealSense SR300[3], (2) Xenics Gobi-640[4], and (3) Seek Thermal Compact

[3] `software.intel.com/en-us/realsense/sr300`
[4] `www.xenics.com`

Pro[5]. Using these devices, presentations are captured in VIS, NIR, and LWIR wavelength bands. Intel's RealSense SR300 camera incorporates two sensors, one for capturing color videos, and another for capturing depth data. It relies on NIR structured light (nominal wavelength: 860nm [38]) to capture depth information. The camera produces the most accurate results in the depth range of 0.2m–1.2m. Therefore, to capture good quality images, the subject should be positioned quite close (0.2m–0.5m) to the camera. Besides color (RGB) imagery, the camera also captures NIR videos. It is important to note that the two cameras (color and NIR) have different fields of view. Color videos of upto full-HD (1920×1080) resolution at a rate of 30 frames per second (fps) can be obtained using this camera. This camera captures NIR images at VGA resolution (640×480) at 30 fps. The Xenics Gobi-640 thermal camera covers a wavelength range of 800nm–1200nm (*i.e.*, long-wave infrared (LWIR)), and captures 16-bit images at VGA resolution at frame-rates up to 50 fps. This camera costs about USD 10000. With this camera, we have used a 18mm $f/1$ lens having a horizontal field of view of 33°. The Compact Pro thermal camera also operates in the LWIR range, and collects thermal images at QVGA resolution at approximately 10 fps. It is designed to work with most mobile phones, and costs about USD 500. For data collection in this work, the three cameras have been deployed in a fixed spatial configuration. We rely on the fixed configuration to determine the mutual spatial calibration of the cameras. Samples images in the different wavelengths collected using the two cameras are shown in Fig. 5. Images in the top row of the figure correspond to a *bona fide* presentation whereas images in the bottom row correspond to a mask-attack presentation. For this study, presentations have been captured using the three cameras simultaneously. The three cameras are mounted in a fixed spatial configuration. The software for triggering the cameras and recording data from the cameras during a data capture session synchronizes the frames from different devices using timestamps associated with the frames. Knowledge of the spatial configuration of the cameras is used to establish the geometric correspondence among the channels. This synchronous collection of data in different spectral bands is key to performing multispectral biometrics, because this is the only way to make sure that a given presentation is captured in different spectral bands. Sample images of *bona fide* presentations are shown in Fig. 5. The images in Fig. 5(a) and (b) have been captured using the Realsense SR300 camera, and show the *bona fide* presentation in visible wavelengths (RGB) and NIR band respectively. Figures 5(c) and (d) show images captured using the Xenics Gobi and the Compact Pro thermal cameras, respectively, illustrating the appearance of a *bona fide* presentation in the LWIR band.

There is an important reason why we have used two different thermal cameras to capture data for this study. A previous study [32] has already established that thermal imagery is highly suited for detecting mask-based PAs. That study used the Xenics Gobi-640 camera, which, while providing imagery of excellent quality, is too expensive to be practical for widespread use in face-PAD solu-

[5] www.thermal.com/compact-series.html

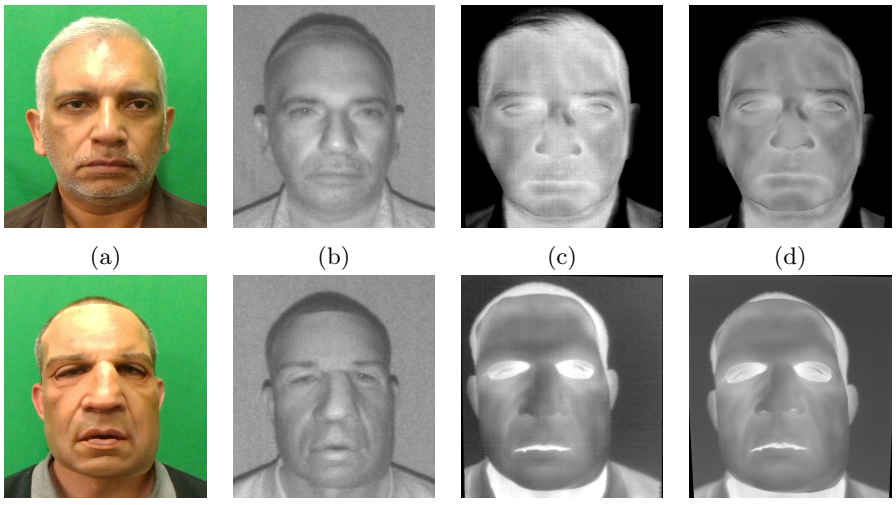


Figure 5: Examples of custom silicone mask-based PAs captured using extended range imaging. (a)–(d) *bona fide* samples; (e)–(h) custom-mask attack presentations. The imaging modalities are, from left to right, VIS, NIR, high-resolution LWIR and low-resolution LWIR, respectively. Note that the reflectance of the mask is similar to that of the *bona fide* presentation in VIS and NIR bands, but significantly different in LWIR images.

tions. Consumer-grade thermal cameras, such as the Compact Pro, albeit of much lower quality, are now available at fairly affordable cost. One of the main goals of our study is to determine whether the face-PAD performance similar to that obtained using the Xenics Gobi camera can be achieved using recently available low cost thermal cameras. For this reason we have also captured LWIR data using the Compact-Pro camera in this study.

3.3 The XCSMAD Dataset

The XCSMAD dataset consists of 240 *bona fide* and 295 PA videos (each $\approx$ 10 s in duration). Each PA video shows a frontal view of a presentation. Since the custom-masks have openings for eyes and mouth, in a PA the attacker can perform actions such as speaking, specific lip movements, and eye-blinking. Some statistics of the XCSMAD and other datasets for 3D-mask based PAs are summarized in Table **??**.

For each presentation, four channels of video data are captured: VIS, NIR (860nm), and two LWIR channels, one using the lower quality (and low cost) Compact Pro thermal camera, and the other using the much more expensive, Xenics-Gobi thermal camera. The data-capture devices are described in Section 3.2.

The four channels are temporally synchronized. We rely on this synchronization for face localization. The MTCNN [39] used for face localization has

been trained for VIS images. Using the relative positions of the imaging devices, an affine-transform is applied to the face location in a given frame in the VIS channel to locate the face region in the corresponding frames in the other spectral-channels.

The experiments reported in this work have been performed using two different protocols: the *grandtest* protocol and the *CV* (cross-validation) protocol. In the grandtest protocol (Tab. 3) the dataset is split into three disjoint sets: one for training (`train`), one for tuning hyper-parameters, or development (`dev`), and one for evaluating the performance of the tuned system (`eval`). From each video, 50 frames have been selected through uniform sampling of the video. The `train` set consists of 86 *bona fide* and 95 PA videos. With 50 frames per video, the `train` set contains 9050 frames. The `dev` set contains 4000 frames from 80 *bona fide* videos and 5750 frames from 115 PA videos, respectively. The `eval` set has 7950 frames from 74 *bona fide* and 85 PA videos. It is important to note that the three sets are subject-wise disjoint, that is, all data from a given subject appears only in one of the three sets. Thus, even though for certain targets we have used multiple custom-masks, no bias is introduced during classifier-training.

For the CV protocols (Tab. 4), we split the data into five non-overlapping partitions, each including roughly 20% of the subjects. Using these five partitions, we create five test-protocols (`cv0`, $\cdots$, `cv4`), such that in each protocol, four of the partitions are used for training, and the remaining one is used for evaluation. The evaluation partition is different in each of the CV protocols. Similar to the grandtest protocol, here we select 50 frames from each video sampled uniformly throughout the video.

Table 3: Grandtest protocol for the *XCSMAD* dataset.

Partition	**# Videos**	**# Frames**	**Split ratio (%)**	**Total Frames**
train-*bona fide*	86	4300	47.52	9050 (34%)
train-attack	95	4750	52.48	
dev-*bona fide*	80	4000	41.03	9750 (36%)
dev-attack	115	5750	58.97	
eval-*bona fide*	74	3700	46.54	7950 (30%)
eval-attack	85	4250	53.46	
Total	535	26750		26750

Table 4: Cross-validation (CV) protocols of *XCSMAD* dataset.

Protocol	**# `train` Videos** [BF, PA]	**# `eval` Videos** [BF, PA]
cv0	409 [182, 227]	126 [58, 68]
cv1	410 [188, 222]	125 [52, 73]
cv2	433 [194, 239]	102 [46, 56]
cv3	454 [202, 252]	081 [38, 43]
cv4	434 [194, 240]	101 [46, 55]

In the remainder of this paper the label *THE-LQ* refers to LWIR data from the Compact Pro camera, and the label *THE-HQ* refers to LWIR data from the Xenics Gobi camera.

4 Proposed Approach

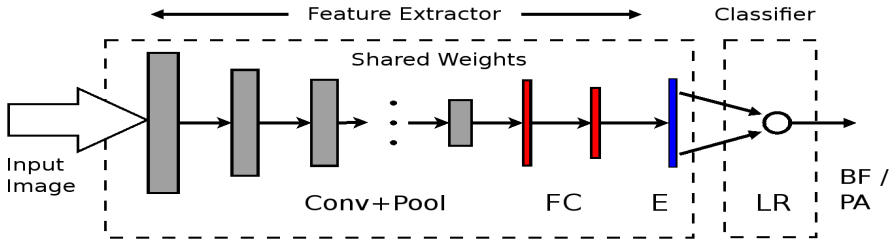


Figure 6: Proposed CNN-based Face-PAD framework.

We propose a CNN-based method to detect custom silicone mask-based PAs. We hypothesize that a CNN pre-trained for face recognition using data from visible spectra can be an efficient *feature extractor* for face-PAD without retraining for transfer learning across different spectral channels. In the context of deep networks, fine-tuning is a common learning process where a network model originally trained for a particular task is partially modified or retrained to perform a similar, but different task. Usually one or more fully connected layers of the CNN are retrained using the specific input data without modifying the lower layers. We hypothesize that CNNs can be used as feature extractors for face-PAD, without explicitly fine-tuning the network for that purpose. We show that *the embeddings extracted from an FR CNN can directly be used to detect silicone mask-based PAs in ER imagery scenarios.* This result is significant not only because it is counter-intuitive, but also because this can lead to simple face-PAD solutions. Let us consider a generic deep CNN consisting of multiple pairs of Convolutional + Pooling layers (`Conv + Pool`) followed by one or more fully connected (`FC`) layers as shown in Fig. 6. The output of pre-final layer of the CNN represents the embeddings (indicated by layer `E` in Fig 6). For an FR CNN, a compact face-representation (or *embedding*) is generated at the pre-final layer from the shared-weights across previous layers. We propose that these embeddings, with the help of an appropriate classifier, can be used to discriminate the *bona fide* and attack presentations. In our experiments, we use a two-class logistic regression (LR) classifier, for its simplicity and efficacy.

To the best of our knowledge, no pre-trained CNN for multi-channel FR or face-PAD is publicly available. Domain adaptation techniques are often employed when the input data and the data used to train the CNN are from different imaging channels or modalities. Typically, transfer learning is used to retrain the upper layers of a pre-trained CNN (layers closer to the classifier-end of the CNN) while keeping the initial (lower) layers frozen. Our preliminary experiments showed that even this form of transfer-learning is not necessary

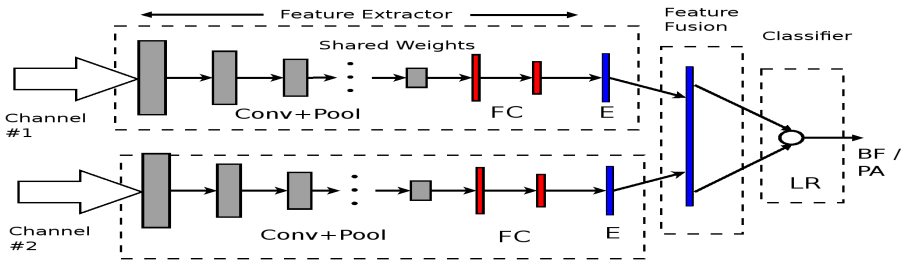


Figure 7: Multispectral feature-fusion face-PAD framework.

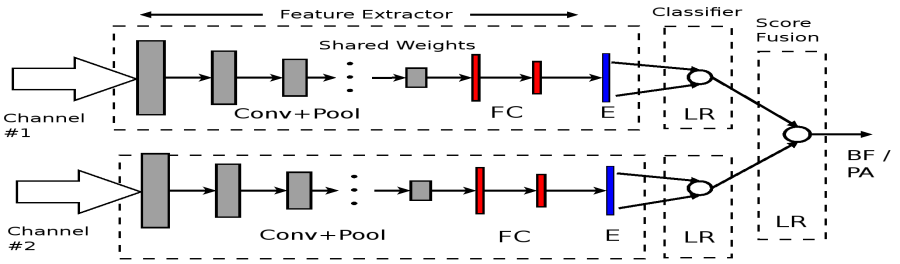


Figure 8: Multispectral score-fusion face-PAD framework.

to adapt a CNN trained for FR tasks to perform face-PAD using ER imagery. Many of the experiments reported here have been performed to verify if this is indeed the case. There are essentially two hypotheses to be validated. First, whether a CNN trained on a VIS images can generate efficient representation of the face from different spectra without any explicit processing. Second, can such a representation be aimed at the detection of presentation attacks—which is different from the original objective of face recognition. We propose to obtain the embeddings of the face images from NIR and LWIR channels using the pre-trained LightCNN. For each spectral band, the embeddings are classified using a two-class LR classifier, trained specifically for that band.

Another aim of this study is to verify whether fusing information from different spectral bands leads to better face-PAD performance compared to single-channel (typically VIS) systems. Therefore, we conduct a series of experiments combining several channels for face-PAD. The complementary information from different channels can be fused at different processing stages. In particular, we have used two different frameworks to explore different multi-channel fusion strategies: *feature fusion* and *score fusion*.

Feature fusion: The features for each channel are extracted separately extracted, and are concatenated in a fixed order to construct a single FV that is input to the classifier (Fig. 7).

Score fusion: Data from each channel is processed separately, and the various classifier-scores are combined using a pre-determined formula, to obtain a single score, that may then be thresholded to generate a decision (Fig. 8).

5 Experiments and Results

Our experiments to evaluate various countermeasures for custom silicone-mask based impersonation attacks are described in this section. As described in Section 3.3, we have collected a new dataset for the work presented in this paper. In Section 5.4 we analyze the vulnerability of the LightCNN based FR system to the custom silicone mask based PAs included in the XCSMAD. This is followed by discussions of our face-PAD experiments presented in several sections.

This is the first study to test face-PAD methods for custom silicone-mask based impersonation attacks. In the first set of experiments, therefore, we establish benchmark performance values of some existing face-PAD approaches (which were originally devised as VIS imagery based countermeasures for 2D PAs). The next logical step is to test these benchmark face-PAD methods on ER imagery. Given that CNN used here operates on single channel input, we employ feature- and score-fusion strategies to combine multiple imaging-modalities in a CNN-based face-PAD.

The term *experiment* here refers to a specific combination of feature extractor and classifier for face-PAD. We refer to each experiment with a label composed of the (<*feature-type*>+ < *classifier*>). For example, **LBP+LR** implies that in the corresponding experiment LBP features were classified using a LR classifier. We begin with a brief description of the experimental setup and performance measures. Before discussing the various experiments, we describe the pre-processing steps used in the various frameworks.

5.1 Overview of Experimental Setup

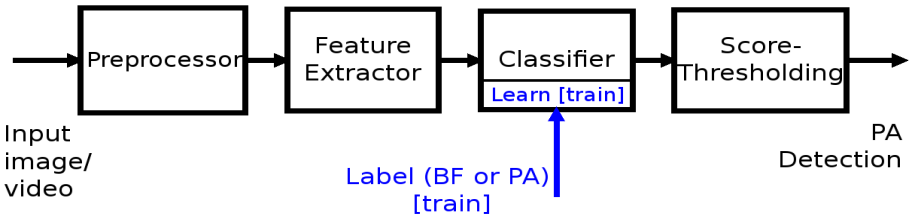


Figure 9: Framework of a single-channel Face-PAD system.

Figure 9 depicts the framework for a single-channel face-PAD system. The input is a presentation in the form of an image from a single channel (VIS, NIR or LWIR). The preprocessing step prepares the input image for feature-extraction. The feature-extractor generates a lower dimensional representation of the input, which is fed to the classifier. In all our frameworks, the face-PAD system functions in one of two phases: training, or evaluation. In the training phase the classifier is trained using the feature-vectors from the `train` set. For every input image the classifier produces a score. The `dev` set is used to determine a score-threshold that best separates the two classes (*bona fide* presentations, and PAs). Here, we have used the equal-error rate (EER) on the

`dev` set to select the score-threshold. The score-threshold τ_{EER} from the `dev` set is then used to classify every presentation in the `eval` set based on the score assigned to the presentation by the classifier. We have used two-class classifiers in all the experiments reported here– hence the use of labels during the training process); but a one-class classifier may also be used in this framework.

In this work we have implemented this framework using two-class classifiers. One-class classifiers may also be used in this framework. Our experiments, however, showed that one-class classifiers produce significantly worse results than their two-class equivalents. Therefore, due to space constraints, we have not discussed one-class classification results in this work.

5.2 Performance Measures

The ISO/IEC 30107-3:2017 standard specified two measures for reporting the performance of a PAD system: APCER (attack presentation classification error rate) and BPCER (*bona fide* presentation classification error rate). APCER is the proportion of presentation attacks (PA) incorrectly classified as *bona fide*. BPCER is the proportion of *bona fide* presentations incorrectly classified as PAs. For ease of comparison, we also provide the average classification error rate (ACER), computed as $\frac{\text{APCER}+\text{BPCER}}{2}$. The vulnerability of a FR system to PAs is reported as the impostor attack presentation match rate (IAPMR), defined as the proportion of PAs accepted as genuine presentations.

5.3 Preprocessing

The input image is pre-processed in two ways: (1) face-localization and resizing, and (2) normalization of radiometric characteristics of the face-region. The pre-processing steps are implemented differently for the various imaging-modalities.
Face-region normalization: For images from the VIS channel we use the Multi-task Cascaded Convolutional Networks (MTCNN) [39] to detect the facial region. Since no pre-trained face-detector is available for NIR and LWIR images, we use information about the spatial configuration of the various cameras to localize the facial region in these two imaging-modalities via appropriate affine-transforms for each camera, using the VIS channel as reference. The cropped facial regions then are resampled to 64×64 pixels.
Radiometric normalization: The feature-extraction stage expects an 8-bit gray-scale image as input. For color images from VIS channel the gray-scale representation is generated by converting the RGB image to a YCbCr representation and then simply considering the Y-channel. The NIR and LWIR images are natively represented as 16-bit gray-scale images. We convert the 16-bit IR images to 8-bit images by first clipping the pixels to range computed adaptively based on the median value in the face-region, and subsequently rescaling the pixel values to the integer-range [0–255].

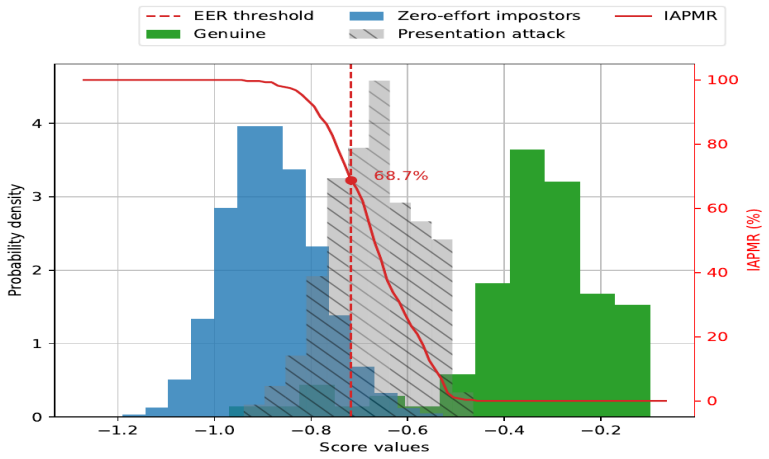


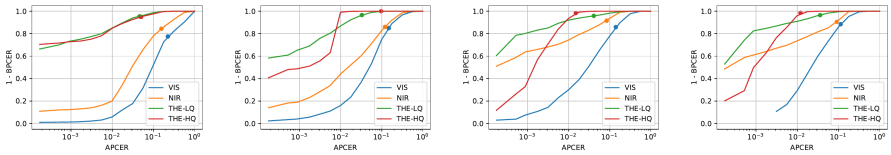
Figure 10: Score-distribution histograms for vulnerability analysis of LightCNN based FR method, using XCSMAD.

5.4 Vulnerability Analysis of LightCNN using XCSMAD

In this section, we describe the vulnerability analysis experiments performed using the XCSMAD. To assess the threat posed by custom-mask PAs in the XCSMAD, we create an experimental protocol similar to the one described in [1]. The *bona fide* presentations include the videos or high quality photos of each subject. For each of the 75 subjects, one *bona fide* presentation is used for enrolment, and the remaining presentations act as probes. Each PA video constitutes an attack, and is used as a probe against the corresponding target-identity. We follow the same preprocessing and feature extraction steps as used for our face-PAD experiments to obtain a 256-dimensional feature vector (FV) per sample. The cosine-similarity is used to compute the matching score between a given probe FV and the enrolment FV of the claimed identity.

The protocol used for the vulnerability analysis experiment produces a total of 13,986 zero-effort impostor (ZEI) probes and 278 custom-mask presentations. First, the match scores of the genuine and ZEI probes are used to select a score-threshold. Here, we have used the *a posteriori* EER to determine the score-threshold for the vulnerability analysis. Next, the match-scores of the mask PA probes are evaluated against this score-threshold to determine the vulnerability of the CNN-FR system to these custom-mask attacks.

For the given protocol and EER threshold, we obtain an IAPMR of 68.71% (the 95% confidence interval for the IAPMR being [62.90, 74.11]). Figure 10 shows the score distributions of each type of presentation. The middle histogram represents the distribution of the PA scores. Note that a majority of the PA scores lies above the EER score-threshold between genuine and ZEI presentation-scores.



(a) PCA+LDA (b) IQM+LR (c) LBP+LR (d) LBP+SVM

Figure 11: ROC of `dev` set for various baseline face-PAD methods. On each curve, the circle indicates corresponding EER-threshold.

5.5 Single Channel Face-PAD

Table 5: Performance of face-PAD methods using the grandtest protocol for individual imaging-modalities in XCSMAD. EER values are determined from the `dev` set, whereas the APCER, BPCER and ACER values are reported for the `eval` set using the score-threshold corresponding to the EER of the `dev` set. The minimum values of ACER of `eval` set obtained for each imaging modality are displayed in bold.

PAD Method	**Imaging Modality**	**EER** % (`dev`)	**APCER** % (`eval`)	**BPCER** % (`eval`)	**ACER** % (`eval`)
PCA+LDA	VIS	22.42	14.71	25.30	20.00
	NIR	15.62	9.50	11.86	10.68
	THE-LQ	4.60	6.19	1.11	3.65
	THE-HQ	5.03	2.66	4.11	3.38
IQM+LR	VIS	14.99	7.18	20.46	13.82
	NIR	12.83	2.73	8.70	5.72
	THE-LQ	3.37	2.85	1.57	2.21
	THE-HQ	0.97	0.00	2.89	1.45
LBP+LR	VIS	14.10	13.60	11.78	12.69
	NIR	8.45	0.56	0.92	**0.74**
	THE-LQ	4.12	4.71	0.76	2.73
	THE-HQ	1.53	0.21	1.08	0.65
LBP+SVM	VIS	11.35	12.05	10.68	11.36
	NIR	9.28	0.45	1.27	0.86
	THE-LQ	3.58	4.07	1.92	2.99
	THE-HQ	1.25	0.14	5.30	2.72
CNN+LR (Proposed)	VIS	1.88	0.00	7.30	**3.65**
	NIR	1.10	1.34	0.59	0.97
	THE-LQ	0.65	0.00	1.05	**0.53**
	THE-HQ	0.25	0.00	0.00	**0.00**

Next, we present four different experiments, each performed on every data-channel individually. These baseline experiments use open source methods for face-PAD. Quantitative results of these experiments are shown in Table 5. The results of each experiment are also presented graphically using receiver operating characteristics (ROC) curves on the `dev` sets. Subsequently, we describe experiments using our proposed approach.

PCA+LDA: This experiment is aimed at understanding whether Principal Component Analysis (PCA) is able to capture differences in *bona fide* and PA images in the various spectral bands used in this work. For each channel, given n preprocessed *bona fide* images of size 64×64, a $4096 \times n$ matrix is constructed. Via PCA of this matrix, we retain the subset of m PCs that accounts for 80% of the variance in the original matrix. This produces a m-dimensional (m-D) subspace ($m \ll 4096$; m is different for each channel.) Thus, a m-D feature-vector can be derived for each input image (*bona fide* or PA) by projecting the preprocessed image on to this subspace. The m-D feature-vectors are then scored using a linear discriminant analysis (LDA) classifier.

Figure 11a shows the ROC curves for the PCA+LDA experiment on the `dev` set. From these plots, we note that data in both LWIR channels produce significantly better results than VIS and NIR data. Recall that the features used in this experiment are simply weighted linear combinations of normalized pixel-values. For VIS data, an ACER of 20.00% (Tab. 5) indicates that the PCA+LDA method generates the wrong decision for one out of every five images.

IQM+LR: In this experiment we have used the same 18 IQMs as used by Costa-Pazo *et al.* [34]. Here, these measures are computed over the gray-scale face-region, to generate a 18-D FV for each input image. A LR classifier is constructed using IQM feature-vectors corresponding to the `train` set, and the score-threshold is determined using the `dev` set.

Figure 11b shows the ROC curves for the `dev` set for this experiment. The plots show that on `dev` set the THE-HQ channel produces near-perfect results over a large range of APCER. We also note that the THE-LQ data (from the low-cost, Compact Pro thermal camera) also leads to better face-PAD performance than NIR and VIS data. It is interesting to note that the low-quality thermal data outperforms the data from high-quality thermal camera at lower values of APCER. The relatively poor results on VIS and NIR data reflect the fact that the appearances of *bona fide* faces and custom silicone masks are quite similar in both imaging domains (see Fig. 5). Therefore, IQM features are not highly discriminative between the two kinds of presentations in VIS and NIR imagery. From the performance measures reported in Table 5 for this experiment, we note that the ACER decreases significantly in NIR and LWIR modalities, from 13.82% for the VIS data to 1.45% for THE-HQ data. A perfect APCER of 0% is obtained for THE-HQ data when the IQM are used as discriminatory features, meaning that the only errors are due to misclassifications of *bona fide* presentations.

LBP+LR: Local binary patterns (LBP) and their variants routinely outper-

form other descriptors in 2D face-PAD experiments [40, 41, 42]. Here we have computed uniform $LBP^{u2}_{8,1}$ codes on normalized face images. The LBP-histogram forms a 59-D FV representing the input image. A LR classifier is then designed to score such FVs.

ROC curves (for the `dev` set) for this experiment are shown in Fig. 11c. From the EER and ACER values for this experiment, given in Table 5, we note again that LBP-histograms are far more discriminative in NIR and thermal channels than for the VIS channel. For the NIR channel, the LBP+LR architecture produces excellent performance on the `eval` set (APCER and BPCER are both smaller than 1%). On `dev` set, however, the NIR channel produces a relatively high EER. Similar to the IQM+LR experiment, in the `dev` set, the data captured from low-quality thermal camera produces the best results at lower ranges of APCER. In the `eval` set, although the THE-HQ data produces the best results, the results from the THE-LQ data are comparable to those of the THE-HQ data (relative to the VIS channel data).

The two PAD systems discussed so far, IQM+LR and LBP+LR, use different features but employ the same classification method. Our results indicate that two systems produce similar results for VIS channel, in terms of EER and ACER, and they both produce significantly better results for the NIR and LWIR data than for the VIS data.

LBP+SVM: Support vector machine (SVM) classifiers have been shown to yield better results for face-PAD in some previous studies [6, 34]. Therefore, in this experiment we use a SVM, with radial basis function (RBF) kernel, to classify the LBP-histogram FVs. From the performance numbers for this experiment (see Table 5), we note that for the `dev` set the RBF-SVM classifier performs similarly to the LR classifier used in the LBP+LR experiment, except for VIS channel where an improvement in EER of nearly 3% is observed. Comparing the ROC curves of this experiment (Fig. 11d) with those shown in Fig. 11c leads us to the same conclusion. The performance numbers for this experiment (Table 5) also confirm the trend seen in the previous two experiments, namely, that the use of NIR and LWIR data brings at least a three- to four-fold improvement in detecting custom-mask based PAs, over the use of the VIS data alone. In this experiment, THE-HQ channel produces a near perfect face-PAD performance on the `dev` set across a large range of APCER (see Fig. 11d). Over the `eval` set, the performance on the THE-HQ channel is worse, as a BPCER of 5.3% is obtained for the computed τ_{EER} score-threshold.

Interestingly, for the NIR-channel, the LBP-based methods produce BPCER (false rejection) values much smaller than for other baseline methods. Also, for both LBP-based methods, the NIR channel yields APCER as well as ACER below 1% on the `eval` set. In our experiments, however, the LBP-histograms are less discriminatory in the low-cost LWIR channel (THE-LQ), as indicated by the relatively high APCER for this channel. Next we propose an approach to PAD for custom silicone mask attacks based on a CNN.

Proposed Approach (CNN+LR): Here, we utilize the 9-layer LightCNN [9] as a feature-extractor (Model taken from: `github.com/AlfredXiangWu/LightCNN`). LightCNN has been developed for FR tasks in several variants (each variant hav-

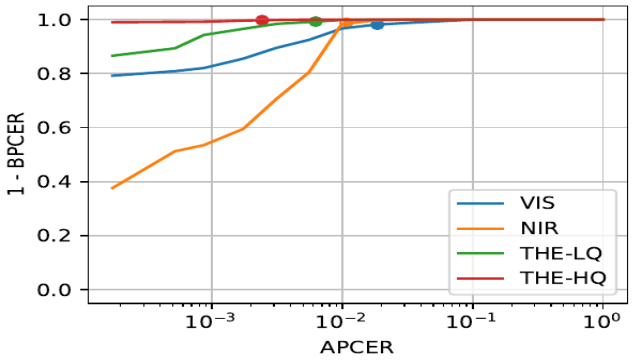


Figure 12: ROC of `dev` set for the proposed method: CNN+LR

ing a different number of layers), and is one of the most accurate FR methods today (achieving $\approx$ 98.8% accuracy [9] on the LFW dataset). In this work we directly use the 256-D embedding produced by the first fully-connected layer of LightCNN (termed `MFM_fc1` by the creators of LightCNN [9]) to characterize face-regions in the XCSMAD data.

The preprocessing step in this experiment generates 128×128 face-cropped gray-scale images as input to the LightCNN network. It is important to note that here the LightCNN pre-trained for FR tasks has not been explicitly adapted for face-PAD. Also, recall that the LightCNN model has been trained using VIS (RGB color) images [9] only. Here, we consider the embeddings from the `MFM_fc1` layer of LightCNN, without nay adaptation of the network to other wavelength-bands (NIR and LWIR), as the FV for a presentation-image. A two-class LR classifier is then trained to score the 256-D FVs.

As reported in Table 5, the proposed method produces a perfect APCER of 0.00% for the three channels- VIS, THE-LQ, and THE-HQ. In other words, for these channels, no PA in the `eval` set is incorrectly classified as *bona fide*. In terms of ACER, for the VIS channel the proposed CNN+LR method outperforms, by a factor of three to four, all the baseline methods previously discussed in this study. Similarly for NIR images in the `eval` set, this method produces lower BPCER (0.6%) than the baseline methods. The ACER of 0.53% for THE-LQ images is the lowest error rate observed for this channel among all experiments. For the THE-HQ channel, the proposed CNN+LR method produces perfect results. One noteworthy observation from this experiment is the performance of this method on the `dev` set. For each imaging channel, the best values of EER are obtained for the proposed method. From the ROC plots over the `dev` set (Fig. 12), we also note that the proposed method produces excellent results for APCER values below 0.01%, regardless the imaging channel.

Finally, we briefly discuss results of proposed approach using the CV protocols described in Table 4. Since these protocols do not explicitly include `dev` sets, we report the *a posteriori* EER values of each of the 5 protocols along with their average. The results are summarized in Table 6. Comparing these results

Table 6: Performance evaluation of the proposed approach on XCSMAD dataset using cross-validation (CV) protocol.

Imaging Modality	**EER** `cv0`	**EER** `cv1`	**EER** `cv2`	**EER** `cv3`	**EER** `cv4`	**EER** `avg`
VIS	1.38	0.66	1.00	0.32	0.87	0.85
NIR	0.11	0.08	0.39	0.57	0.26	0.28
THE-LQ	0.35	0.05	1.74	0.00	0.00	0.43
THE-HQ	0.00	0.08	0.00	0.00	0.00	0.02

with those for the CNN+LR experiment on the grandtest protocol (bottom of the EER column in Table 5), we note that the error-rates in the CV protocols are lower than in the grandtest protocol, for all four data-channels. The improved performances in the CV protocols reflect the fact that in these protocols the size of training data for the LR classifier is significantly larger than the size of the `train` set in the grandtest protocol.

Table 6 shows the results of the proposed CNN+LR method using the CV protocols specified in Section 3.3. We note that, for each imaging channel, the CV results are quite similar to those obtained using the grandtest protocol. (Similar CV experiments were also performed for the four baseline methods. Again, the results with the various CV partitions were similar to those listed in Table 5. Due to space constraints, we have not included details of the CV results here. The source-code for all our experiments, including the CV experiments, is publicly available. We invite the reader to verify the CV results using the published code.)

Figure 13 shows some of the presentations classified incorrectly by the proposed method. The two left columns show examples of misclassified *bona fide* samples, whereas the two right columns show examples of successful PAs, in different imaging modalities. When using CNN embeddings, it is not possible to pinpoint the reason for a misclassification. There does not seem to be any obvious reason for false-rejections in neither the VIS images (Fig. 13 (a), (b)), nor the NIR images (Fig. 13 (e), (f)). For the false-accepts, shown in the two rightmost columns, we note that the images shown here do present realistic facial features, both in NIR and in THE-LQ.

5.6 Fusion Experiments

The next logical step is to explore multispectral fusion strategies for face-PAD. In this section we present results of several pair-wise multispectral fusion experiments. The single-channel experiments (Section 5.5) showed that the THE-HQ data, captured using a high quality but expensive Xenics Gobi thermal camera is not indispensable for reliable detection of custom-silicone mask based PAs. Data

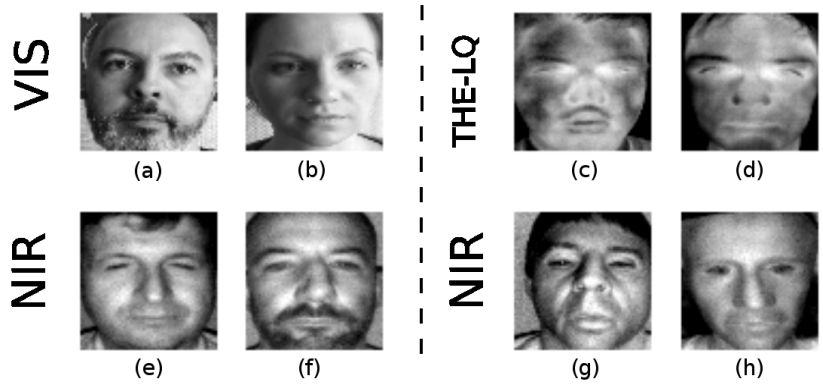


Figure 13: Examples of presentations misclassified by the proposed approach. (a) & (b): misclassified *bona fide* from VIS channel. (c) & (d): misclassified *bona fide* from THE-LQ channel. (e) & (f): misclassified *bona fide* samples from NIR channel. (g) & (h) misclassified PAs from NIR channels. All images are preprocessed face regions.

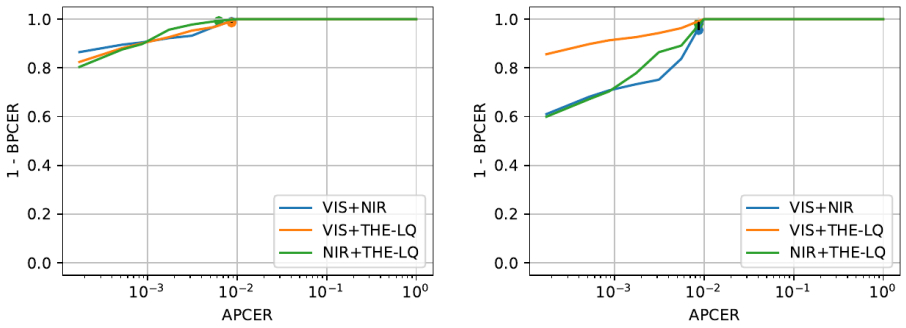
from low-cost thermal cameras– Seek Thermal's Compact Pro, in our case– can be used to achieve a comparable performance (difference in error rates is often within 1%). Therefore, in the set of multispectral fusion experiments discussed in this section, we have not included the THE-HQ data.

Both fusion strategies– feature fusion and score fusion– have been tested on different combinations of spectral channels. Performance-metrics for the fusion experiments are listed in Table 7. For each set of experiments, we also present the corresponding ROC curves in the discussion below.

Feature Fusion: Given the *embeddings* of two normalized face-images (outputs of the preprocessing stage) of the same presentation from different channels, these FVs of the two images are stacked to obtain a single concatenated FV of length $k = k_1 + k_2$, where k_1 and k_2 are the lengths of feature vectors of first and second images, respectively. In our case, both FVs are of the same length since we employ the same CNN as feature extractor. A two-class LR classifier is designed to score these k-D FVs. We have considered the following three spectral combinations: VIS + NIR, VIS + THE-LQ and NIR + THE-LQ. A bespoke LR classifier is trained for each spectral-pair.

Figure 14a shows the ROC plots of `dev` set for feature-fusion experiments. Quantitative performance measures are reported in Table 7. In the VIS+NIR feature fusion experiment the EER and ACER obtained by combining data from the two channels are better than the corresponding single-channel CNN+LR experiment results (compare with Table 5). Note that this fusion retains the perfect APCER observed in the single-channel experiment using VIS data.

When data from VIS and THE-LQ channels are fused, an ACER of 2.63% is obtained on the `eval` set, which is lower than the ACER values achieved by the proposed CNN+LR method for single-channel experiments of VIS images. However, the THE-LQ channel produced a better BPCER when used separately than fusing with VIS channel in this experiment.



(a) Feature fusion

(b) Score fusion

Figure 14: ROC of `dev` set for the proposed multi-channel fusion face-PAD methods.

Table 7: Performance Evaluation of Fusion face-PAD experiments on XCSMAD dataset. The values reported are percentages. The minimum values of ACER of `eval` set obtained for each pair of imaging-modalities are displayed in bold.

Fusion Method	**Channels**	**EER** (`dev`)	**APCER** (`eval`)	**BPCER** (`eval`)	**ACER** (`eval`)
Feature fusion	VIS + NIR	0.87	0.00	1.40	**0.70**
	VIS + THE-LQ	0.87	0.00	5.27	2.63
	NIR + THE-LQ	0.63	0.00	0.62	0.31
Score fusion	VIS + NIR	0.87	0.00	2.43	1.22
	VIS + THE-LQ	0.87	0.00	2.11	**1.05**
	NIR + THE-LQ	0.87	0.00	0.40	**0.20**

In general, we can conclude that combining the VIS channel with other wavelength bands (NIR or LWIR) leads to a higher detection accuracy (than when only VIS data is used), for custom-silicone mask PAs. This makes a strong case for using ER imagery for face-PAD. Combining the NIR and THE-LQ images, we observe an improvement of nearly 40% in ACER compared to the corresponding single-channel results using CNN+LR method. The perfect APCER of 0% is achieved in all feature-fusion experiments while lowering the average error rates (ACER). These experiments show that feature-fusion can be effective in improving PAD accuracy for custom-silicone masks when multiple imaging channels are available.

Score Fusion: The framework for our proposed score fusion scheme is depicted in Fig. 8. For score based fusion of multispectral data, a separate face-PAD system is implemented for each spectral channel. Given a presentation, the n

scores thus obtained, for n imaging channels, are then combined using an *a priori* fusion scheme. The fusion scheme may be rule-based, such as taking the minimum or maximum of the available scores, or may be derived algorithmically, by training a classifier in a data-driven fashion. The final fused score may then be thresholded appropriately, to arrive at a decision. Here we train a two-class LR classifier to combine the three single-channel LightCNN scores into a single fused score. The LR classifier is trained on the scores computed over the `train` set, and the EER score-threshold, τ_{EER}, is determined using the `dev` set. The ROC curves for this experiment are plotted in Fig. 14b. In these plots, ROC curves for three channel-fusion experiments are shown, namely, VIS + NIR, VIS + THE-LQ, and NIR + THE-LQ. For all three experiments, score-fusion leads to a APCER of 0.0%, as observed in the results of proposed CNN+LR method for individual VIS and THE-LQ channels. It is also worth noting that a consistent value of EER is obtained on the `dev` set for all three fusion pairs. For score-fusion of VIS and NIR channels, this EER value (0.87%) is lower than the EER obtained using the single channel CNN+LR method. Note that the ROC plots of the CNN+LR fusion scheme are practically the same for VIS + NIR, VIS + THE-LQ, and NIR + THE-LQ channel combinations in the `dev` set for APCER below 0.01%. Consider the fusion of NIR + THE-LQ channels: The performance metrics presented in Table 7 indicates that ACER for combination is less than 0.3% compared to individual channel experiments. In the proposed CNN+LR method, the NIR channel resulted in a APCER of 1.34%, while THE-LQ channel produced a perfect APCER. On the other hand, NIR channel produced nearly half value of BPCER as compared to the THE-LQ channel. Our experiments show that the score-fusion of these channels lowers the BPCER by nearly 30% of the BPCER of individual NIR channel, while maintaining the APCER at 0%.

The single-channel experiments (Sec. 5.5), confirm findings in previous works ([32]) that the LWIR modality (both THE-LQ and THE-HQ data) produces the best PAD results for 3D custom silicone masks. Analyzing the results of the spectral fusion experiments, we note that the inclusion of thermal data, for both scenarios—feature fusion and score fusion leads to better face-PAD performance for the class of PAs considered in this study.

6 Conclusion

We present the first data-driven study on detecting PAs mounted using custom-made silicone masks[6]. For this study, Nimba Creations Ltd., a special-effects company, has constructed very realistic silicone masks for 17 target-subjects, based on pictures and 3D scans provided by us. State of the art CNN-FR systems have been shown to be highly vulnerable to PAs made using such masks [1, 2]. In this work we present a new method for detecting impersonation PAs made using custom silicone masks. The proposed method uses a CNN to extract

[6]Python code for all experiments mentioned in this work is available at: `https://gitlab.idiap.ch/bob/bob.paper.xcsmad_facepad`

a feature-vector (FV) for every input image. The FV is then classified using a logistic-regression (LR) classifier. We compare the performance of the proposed CNN+LR method with several commonly used baseline face-PAD techniques. These baseline methods have been developed mainly for 2D PAs such as print-attacks or replay-attacks. Here we have explored the efficacy of these techniques for the specific case of custom 3D silicone masks.

Another significant aspect of this study has been the use of extended-range (ER) imagery for detecting custom silicone-mask based PAs. Specifically, we have used data captured in several spectral wavelength-bands: visible-light (VIS), near-infrared (NIR) and thermal (or, long-wave infrared (LWIR)). Our experiments show that the baseline methods, which have been developed with VIS imagery in mind, produce better results in the NIR and LWIR domains. Other research works have previously demonstrated that imagery in longer wavelengths (specifically SWIR and LWIR) are highly suited for constructing countermeasures for mask based attacks. These earlier studies, however, relied on very expensive imaging systems, typically costing more than USD 10,000. One goal of this study has been to explore whether a new crop of low-cost consumer-grade thermal cameras may be applicable in this context. Our single-channel experiments show that data from a low-cost thermal camera, such as the Compact Pro from Seek Thermal, can lead to PAD performance comparable to that obtained using the Xenics Gobi thermal camera, which is 20 times more expensive.

The proposed custom silicone mask PAD method used embeddings extracted from a 9-layer LightCNN [9] as FVs. It is important to note that the LightCNN has been trained for FR tasks, only on VIS images. It is one of the best performing FR methods available today. Our experiments lead to a very counter-intuitive observation that some features that perform exceedingly well for FR, can also lead to near-perfect performance for custom silicone mask PAD. Our single-channel experiments show that the CNN+LR method performs better than the baseline methods for every imaging channel. For VIS and two thermal channels, the proposed method results in the perfect APCER of 0.0%. The spectral fusion experiments with the proposed CNN+LR method show that both feature-fusion and score-fusion approaches outperform all other methods discussed in this work, consistently producing extremely low BPCER values while maintaining zero APCER. This is the first work to demonstrate that not only can the embeddings from LightCNN (trained for FR tasks using VIS images) be used for face-PAD applications using VIS data, but the same embeddings, without explicit domain adaptation, also lead to very high accuracy face-PAD approaches in other imaging-modalities, specifically in NIR and LWIR spectral bands.

In future work we will examine ways of fusing ER-imagery based PAD for custom-masks with face-PAD approaches for other classes of PAs to devise a unified face-PAD strategy. We will explore approaches for combining such a unified PAD system with FR systems, to construct a trustworthy face-authentication system.

Acknowledgments

This work has been supported by the European H2020-ICT project TeSLA (grant agreement no. 688520), and by the Swiss Center for Biometrics Research and Testing. We are grateful for the assistance of our colleagues, Z. Mostaani, G. Clivaz, and D. Geissbuhler, who developed the imaging platform, within the framework of the IARPA ODIN project, to collect data for this study. We also thank Mr. Thomas Lauten of Nimba Creations Ltd. who constructed the custom-silicone masks used here.

References

[1] S. Bhattacharjee, A. Mohammadi, and S. Marcel, "Spoofing Deep Face Recognition with Custom Silicone Masks," in *Proc. IEEE Conf. on Biometrics: Theory, Applications and Systems*, Los Angeles, USA, 10 2018.

[2] R. Raghavendra, S. Venkatesh, K. B. Raja, S. Bhattacharjee, P. Wasnik, S. Marcel, and C. Busch, "Custom silicone face masks - vulnerability of commercial face recognition systems & presentation attack detection," in *Proc. 7th IAPR/IEEE Intl. Workshop on Biometrics and Forensics (IWBF)*, May 2019.

[3] R. Ramachandra *et al.*, "On the Vulnerability of Extended Multispectral Face Recognition Systems Towards Presentation Attacks," in *Proceedings of IEEE International Conference on Identity, Security and Behavior Analysis*, New Delhi, 2017, pp. 1–8.

[4] A. Mohammadi, S. Bhattacharjee, and S. Marcel, "Deeply Vulnerable–A Study of The Robustness of Face Recognition Methods to Presentation Attacks," *IET Biometrics*, vol. 7, no. 1, pp. 15–26, 2018.

[5] S. Bhattacharjee *et al.*, "Recent Advances in Face Presentation Attack Detection," in *Handbook of Biometric Anti-Spoofing*, S. Marcel, M. S. Nixon, J. Fierrez, and N. Evans, Eds. Springer Nature Switzerland AG, 01 2019, pp. 207–228.

[6] R. Ramachandra and C. Busch, "Presentation attack detection methods for face recognition systems: A comprehensive survey," *ACM Computing Surveys*, vol. 50, no. 1, pp. 8:1–8:37, 2017.

[7] N. Erdogmus and S. Marcel, "Spoofing in 2D Face Recognition With 3D Masks and Anti-spoofing With Kinect," in *Proc. IEEE Intl. Conf. BTAS*, Washington D.C., 2013.

[8] S. Liu *et al.*, "A 3D Mask Face Anti-Spoofing Database with Real World Variations," in *Proceedings of IEEE Conf. on Computer Vision and Pattern Recognition Workshops*, Las Vegas, NV, USA, June 2016, pp. 1551–1557.

[9] X. Wu *et al.*, "A Light CNN for Deep Face Representation With Noisy Labels," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 11, pp. 2884–2896, Nov 2018.

[10] A. Agarwal, D. Yadav, N. Kohli, R. Singh, M. Vatsa, and A. Noore, "Face Presentation Attack with Latex Masks in Multispectral Videos," in *Proceedings of IEEE Conf. on Computer Vision and Pattern Recognition Workshops*, July 2017, pp. 275–283.

[11] I. Manjani *et al.*, "Detecting Silicone Mask-Based Presentation Attack via Deep Dictionary Learning," *IEEE Trans. Info. Forensics and Security*, vol. 12, no. 7, pp. 1713–1723, 2017.

[12] N. Kose and J. Dugelay, "Countermeasure for the Protection of Face Recognition Systems Against Mask Attacks," in *Proceedings of IEEE Intl Conf. on Automatic Face and Gesture Recognition*, 2013.

[13] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[14] S. Lina and L. Ramavel, "Masquerade Attack Detection by Analyzing Local and Global Features in Face Recognition System," *International Journal of Applied Engineering Research*, vol. 10, pp. 162–166, 2015.

[15] A. Agarwal, R. Singh, and M. Vatsa, "Face Anti-Spoofng Using Haralick Features," in *Proceedings of the IEEE International Conf. on Biometrics: Theory, Applications, and Systems*, Niagara Falls, NY, USA, 09 2016, pp. 1–6.

[16] X. Li *et al.*, "Generalized Face Anti-Spoofing by Detecting Pulse from Face Videos," in *Intl. Conf. on Pattern Recognition (ICPR)*, 2016, pp. 4244–4249.

[17] S. Liu, X. Lan, and P. Yuen, "Remote Photoplethysmography Correspondence Feature for 3D Mask Face Presentation Attack Detection," in *Proc. European Conf. on Computer Vision*, 2018.

[18] E. Nowara, A. Sabharwal, and A. Veeraraghavan, "Ppgsecure: Biometric presentation attack detection using photopletysmograms," in *IEEE International Conference on Automatic Face Gesture Recognition*, May 2017, pp. 56–62.

[19] G. Heusch and S. Marcel, "Pulse-Based Features for Face Presentation Attack Detection," in *Proc. IEEE Conf. on Biometrics: Theory, Applications and Systems (BTAS) Special Session on Image and Video Forensics in Biometrics*, 2018.

[20] H. Muckenhirn *et al.*, "Long-Term Spectral Statistics for Voice Presentation Attack Detection," *IEEE/ACM Trans. Audio, Speech and Language Processing*, vol. 25, no. 11, pp. 2098–2111, 2017.

[21] R. Ramachandra *et al.*, "Extended Multispectral Face Presentation Attack Detection: An Approach Based on Fusing Information From Individual Spectral Bands," in *Proc. Intl Conf on Information Fusion*, 2017.

[22] L. Li *et al.*, "An Original Face Anti-Spoofing Approach Using Partial Convolutional Neural Network," in *International Conference on Image Processing Theory, Tools and Applications*, 2016.

[23] K. Patel, H. Han, and A. Jain, "Cross-Database Face Antispoofing with Robust Feature Representation," in *Biometric Recognition*. Cham: Springer International Publishing, 2016, pp. 611–619.

[24] T. Nguyen *et al.*, "Combining Deep and Handcrafted Image Features for Presentation Attack Detection in Face Recognition Systems Using Visible-Light Camera Sensors," in *Sensors*, vol. 18, Feb 2018.

[25] L. Feng *et al.*, "Integration of Image Quality and Motion Cues for Face Anti-spoofing: A Neural Network Approach," *Journal of Visual Communication and Image Representation*, vol. 38, pp. 451–460, 2016.

[26] Y. Rehman, L. Po, and M. Liu, "LiveNet: Improving Features Generalization for Face Liveness Detection Using Convolution Neural Networks," *Expert Systems with Applications*, vol. 108, pp. 159–169, 2018.

[27] H. Li *et al.*, "Learning Generalized Deep Feature Representation for Face Anti-Spoofing," *IEEE Trans. on Information Forensics and Security*, vol. 13, no. 10, pp. 2639–2652, 2018.

[28] Z. Sun, L. Sun, and Q. Li, "Investigation in Spatial-Temporal Domain for Face Spoof Detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 04 2018, pp. 1538–1542.

[29] Y. Kanzawa, Y. Kimura, and T. Naito, "Human Skin Detection by Visible and Near-Infrared Imaging," in *Proc. IAPR Conf. on Machine Vision Applications*. Nara, Japan: IEEE, 2011.

[30] T. Bourlai *et al.*, "On Designing a SWIR Multi-Wavelength Facial-Based Acquisition System," in *Proceedings of SPIE: Infrared Technology and Applications*, vol. 8353, 04 2012.

[31] H. Steiner, A. Kolb, and N. Jung, "Reliable Face Anti-Spoofing Using Multispectral SWIR Imaging," in *Proc. of Intl Conf. on Biometrics*, 2016, pp. 1–8.

[32] S. Bhattacharjee and S. Marcel, "What You Can't See Can Help You – Extended Range Imaging for 3D-Mask Presentation Attacks," in *Proc. Intl. Conf. of Biometrics Special Interest Group*, 2017, pp. 1–8.

[33] O. Parkhi, A. Vedaldi, and A. Zisserman, "Deep Face Recognition," in *British Machine Vision Conference*, 2015.

[34] A. Costa-Pazo *et al.*, "The Replay-Mobile Face Presentation-Attack Database," in *Proc. Intl Conf. of the Biometrics Special Interest Group*, 2016.

[35] A. Gretton *et al.*, "A Kernel Two-Sample Test," *Jrnl. Machine Learning Research*, vol. 13, no. 3, pp. 723–773, 2012.

[36] K. Muandet, D. Balduzzi, and B. Schölkopf, "Domain Generalization via Invariant Feature Representation," in *Proceedings of the International Conference on International Conference on Machine Learning*. JMLR.org, 2013, pp. 10–18.

[37] R. Shao, X. Lan, and P. C. Yuen, "Deep Convolutional Dynamic Texture Learning with Adaptive Channel-Discriminability for 3D Mask Face Anti-Spoofing," in *IEEE International Joint Conference on Biometrics (IJCB)*, Oct 2017, pp. 748 – 755.

[38] *Intel RealSense Camera SR300 Product Datasheet*, 2016, www.mouser.com/pdfdocs/intel_realsense_camera_sr300.pdf, Stand: 01.06.2016.

[39] K. Zhang *et al.*, "Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[40] Z. Boulkenafet *et al.*, "A Competition on Generalized Software-Based Face Presentation Attack Detection in Mobile Scenarios," in *Proc. IEEE Intl Joint Conf. on Biometrics*, Oct 2017, pp. 688–696.

[41] I. Chingovska, A. Anjos, and S. Marcel, "On The Effectiveness of Local Binary Patterns in Face Anti-Spoofing," in *Proc. Intl Conf. of Biometrics Special Interest Group*, 2012.

[42] J. Määttä, A. Hadid, and M. Pietikäinen, "Face Spoofing Detection from Single Images using Micro-Texture Analysis," in *Proc. Intl Joint Conf. on Biometrics*, 2011, pp. 1–7.