Multi-channel Face Presentation Attack Detection Using Deep Learning

Anjith George and Sébastien Marcel

Abstract Face recognition has emerged as a widely used biometric modality. However, its vulnerability to presentation attacks remains a significant security threat. Although Presentation Attack Detection (PAD) methods attempt to remedy this problem, often they fail in generalizing to unseen attacks and environments. As the quality of presentation attack instruments improves over time, achieving reliable PA detection using only visual spectra remains a major challenge. We argue that multi-channel systems could help solve this problem. In this chapter, we first present an approach based on a multi-channel convolutional neural network for the detection of presentation attacks. We further extend this approach to a one-class classifier framework by introducing a novel loss function that forces the network to learn a compact embedding for the *bonafide* class while being far from the representation of attacks. The proposed framework introduces a novel way to learn a robust PAD system from *bonafide* and available (known) attack classes. The superior performance in unseen attack samples in publicly available multi-channel PAD database *WMCA* shows the effectiveness of the proposed approach. Software, data, and protocols for reproducing the results are made publicly available.

1 Introduction

Biometrics provides a secure and convenient means for access control. Facial biometrics is one of the most convenient modalities for biometric authentication due to its non-intrusive nature. Even though facial recognition systems

A. George and S. Marcel are with
Idiap Research Institute,
Centre du Parc, Rue Marconi 19, CH - 1920,
Martigny, Switzerland.
e-mail: (anjith.george,sebastien.marcel)@idiap.ch

achieve human performance in identifying people in many difficult [32] data sets, most facial recognition systems are still vulnerable to presentation attacks (PA), also known as spoofing [1] [41, 42], [29]. Simply showing a printed photo to an unprotected facial recognition system might be enough to fool the system [2]. Vulnerability to presentation attacks limits the reliable deployment of such systems for applications in unsupervised conditions.

According to the ISO [29] standard, a presentation attack is defined as:

> A presentation to the biometric data capture subsystem with the goal of interfering with the operation of the biometric system.

Presentation attacks include both 'impersonation' and 'obfuscation' of identity. Impersonation refers to attacks in which the attacker wants to be recognized as a different person, while, in 'obfuscation' attacks, the goal is to hide the identity of the attacker. The biometric characteristic or object used in a presentation attack is known as a presentation attack instrument (PAI).

Presentation attack refers to an attack using an instrument with the intention to affect the normal operation of the biometric system. Often, features such as color, texture [8], [39], motion [2], and physiological cues [55], [28] and CNN based methods [20] are used for detection of attacks like 2D prints and replays. However, detection of sophisticated attacks like 3D masks and partial attacks are challenging and poses a serious threat to the reliability of face recognition systems. Most of the presentation attack detection (PAD) methods available in prevailing literature try to solve the problem for a limited number of presentation attack instruments and on visible spectrum

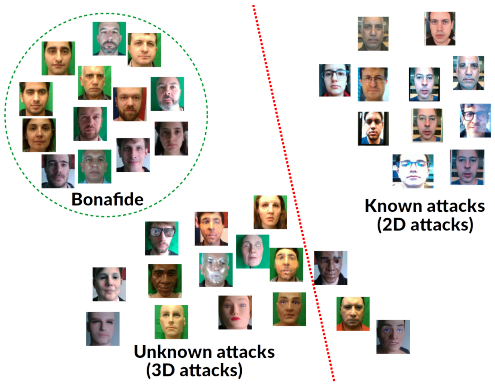


Fig. 1 Illustration of the embedding space with known and unknown attack classes. The red dotted line shows the learned decision boundary when only *bonafide* and known attacks are present in the training set, this results in misclassification of unknown attacks. If a decision boundary of the *bonafide* class (green dotted lines) is learned, known and unknown attacks can be classified correctly.

[1] The term spoofing should be deprecated in favor of presentation attacks to comply with ISO standards.

images [42]. Though some success has been achieved in addressing 2D presentation attacks, performance of the algorithms in realistic 3D masks and other kinds of attacks is poor. With the increase in quality of attack instruments, it becomes harder to discriminate between *bonafide* and PAs in the visible spectrum alone. Moreover, considering a real-world situation with a wide variety of 2D, 3D, and partial attacks, PAD in visual spectra alone is challenging and inadequate for security-critical applications. Partial attacks refer to attacks where the attack instrument covers only a part of the face. These attacks are much harder to detect as they appear similar to *bonafide* in most of the face regions, and they can fool holistic liveliness detection systems easily. Multi-channel methods have been proposed as an alternative [54], [57], [5], [23, 44, 21, 26, 45, 22], since they use complementary information from different channels to improve the discrimination between *bonafide* and attacks. In the multi-channel scenario, the additional channels used can be any modality which can provide complementary representation such as depth, infrared, and thermal channels. Multi-channel PAD approaches are more promising in the context of a wide variety of attacks since they make PAD systems harder to fool.

Even with the use of multiple channels, one of the main issues with PAD is its poor generalization to unseen attacks [23]. This is particularly important, since at the time of developing a PAD system, anticipating all possible attacks is impossible. Malicious attackers can always come up with new attacks to fool the PAD systems. In such situations, PAD systems which are robust against unseen attacks are of paramount importance. Moreover, while it is comparatively easy to collect data for attacks like 2D prints and replays, making replicas of challenging presentation attack instruments (PAI) like silicone mask are often very costly [6] and resource-intensive. In this context, it will be ideal to have a framework which can be trained with *bonafide* alone, or with a combination of *bonafide* and easy to manufacture PAIs.

In real-world scenarios, it can be assumed that all presentation attacks are unseen, as it is not possible to foretell all the variations a PAD system could encounter a priori. A toy example of the decision boundary in an unseen attack scenario is illustrated in Fig. 1. Performances in typical PAD databases may not be representative of the performance of a PAD system in real-world conditions. This necessitates the PAD algorithms to be robust against unseen attacks. Since it is easy (in effort and cost) to collect data from more straightforward attacks compared to complex PAIs, we try to learn the representation leveraging the information from PA classes which are available at the training stage (while not over-fitting on the available attacks). To achieves this, we propose a one-class classifier based framework, where the feature representation is learned with a CNN to have discriminative properties. The core of the framework is a multi-channel CNN trained to learn the embedding using a specific loss function. The Multi-Channel Convolutional Neural Network (MC-CNN) architecture efficiently combines multi-channel information for robust detection of presentation attacks. The network uses

a pre-trained LightCNN model as the base network, which obviates the requirement to train the framework from scratch. In MC-CNN only low-level LightCNN features across multiple channels are re-trained, while high-level layers of pre-trained LightCNN remain unchanged. In combination with the new loss function, the network aims at learning a compact representation for the *bonafide* class while leveraging the discriminative information for PAD task.

The source code and protocols to reproduce the results are made available publicly and are accessible at the following link [2].

The rest of the chapter is organized as follows. Section 2 describes the related work with a particular focus on unseen attack detection. Section 3 outlines the proposed framework. Extensive evaluations, comparison with baseline methods, and ablation studies are shown in section 4. Section 5 discusses the importance of the results, and Section 6 presents the conclusions.

2 Related work

Most of the work related to face presentation attack detection addresses detection of 2D attacks, specifically print and 2D replay attacks. A brief review of recent PAD methods is given in this section.

2.1 Feature based approaches for face PAD

For PAD using visible spectrum images, several methods such as detecting motion patterns [2], color texture and histogram based methods in different color spaces, and variants of Local Binary Patterns (LBP) in grayscale [8] and color images [9], [39] have shown good performance. Image quality based features [18] is one of the successful methods available in prevailing literature. Methods identifying moiré patterns [49], and image distortion analysis [59], use the alteration of the images due to the replay artifacts. Most of these methods treat PAD as a binary classification problem which may not generalize well for unseen attacks [46].

Chingovska *et al.* [10] studied the amount of client-specific information present in features used for PAD. They used this information to build client-specific PAD methods. Their method showed a 50% relative improvement and better performance in unseen attack scenarios.

Arashloo *et al.* [3] proposed a new evaluation scheme for unseen attacks. Authors have tested several combinations of binary classifiers and one class classifiers. The performance of one class classifiers was better than binary

[2] Source code: `https://gitlab.idiap.ch/bob/bob.paper.oneclass_mccnn_2019`

classifiers in the unseen attack scenario. BSIF-TOP was found successful in both one class and two class scenarios. However, in cross-dataset evaluations, image quality features were more useful. Nikisins *et al.* [46] proposed a similar one class classification framework using one class Gaussian Mixture Models (GMM). In the feature extraction stage they used a combination of Image Quality Measures (IQM). The experimental part involved an aggregated database consisting of replay attack [9], replay mobile [11], and MSU-MFSD [59] datasets.

Heusch and Marcel [27] recently proposed a method for using features derived from remote photoplethysmography (rPPG). They used the long term spectral statistics (LTSS) of pulse signals obtained from available methods for rPPG extraction. The LTSS features were combined with SVM for PA detection. Their approach obtained better performance than state of the art methods using rPPG in four publicaly available databases.

2.2 CNN based approaches for face PAD

Recently, several authors have reported good performance in PAD using convolutional neural networks (CNN). Gan *et al.* [19] proposed a 3D CNN based approach, which utilized the spatial and temporal features of the video. The proposed approach achieved good results in the case of 2D attacks, prints, and videos. Yang *et al.* [64] proposed a deep CNN architecture for PAD. A preprocessing stage including face detection and face landmark detection is used before feeding the images to the CNN. Once the CNN is trained, the feature representation obtained from CNN is used to train a SVM classifier and used for final PAD task. Boulkenafet *et al.* [7] summarized the performance of the competition on mobile face PAD. The objective was to evaluate the performance of the algorithms under real-world conditions such as unseen sensors, different illumination, and presentation attack instruments. In most of the cases, texture features extracted from color channels performed the best. Li *et al.* [34] proposed a 3D CNN architecture, which utilizes both spatial and temporal nature of videos. The network was first trained after data augmentation with a cross-entropy loss, and then with a specially designed generalization loss, which acts as a regularization factor. The Maximum Mean Discrepancy (MMD) distance among different domains is minimized to improve the generalization property.

There are several works involving various auxiliary information in the CNN training process, mostly focusing on the detection of 2D attacks. Authors use either 2D or 3D CNNs. The main problem of CNN based approaches mentioned above, is the lack of training data, which is usually required to train a network from scratch. One broadly used solution is fine-tuning, rather than a complete training, of the networks trained for face-recognition, or image classification tasks. Another issue is poor generalization in cross-database,

and unseen attacks tests. To circumvent these issues, some researchers have proposed methods to train a CNN using auxiliary tasks, which is shown to improve generalization properties. These approaches are discussed below.

Liu *et al.* [36] presented a novel method for PAD with auxiliary supervision. Instead of training a network end-to-end directly for PAD task, they used CNN-RNN model to estimate the depth with pixel-wise supervision and estimate remote photoplethysmography (rPPG) with sequence-wise supervision. The estimated rPPG and depth were used for PAD task. The addition of the auxiliary task improved the generalization capability.

Atoum *et al.* [4] proposed a two-stream CNN for 2D presentation attack detection by combining a patch-based model and holistic depth maps. For the patch-based model, an end-to-end CNN was trained. In the depth estimation, a fully convolutional network was trained using entire face image. The generated depth map was converted to feature vector by finding the mean values in the $N \times N$ grid. The final PAD score was obtained by fusing the scores from the patch and depth CNNs.

Shao *et al.* [56] proposed a deep convolutional network-based architecture for 3D mask PAD. They tried to capture the subtle differences in facial dynamics using the CNN. Feature maps obtained from the convolutional layer of a pre-trained VGG network was used to extract features in each channel. Optical flow was estimated using the motion constraint equation in each channel. Further, the dynamic texture was learned using the data from different channels. The proposed approach achieved an AUC (Area Under Curve) score of 99.99% in 3DMAD dataset.

2.3 One class models for face PAD

Most of these methods handle the PAD problem as binary classification, which results in classifiers over-fitting to the known attacks resulting in poor generalization to unseen attacks. We focus the further discussion on the detection of unseen attacks. However, it is imperative that methods working for unseen attacks must perform accurately for known attacks as well. One naive solution for such a task is one-class classifiers (OCC). OCC provides a straightforward way of handling the unseen attack scenario by modeling the distribution of the *bonafide* class alone.

Arashloo *et al.*[3] and Nikisins *et al.* [46] have shown the effectiveness of one class methods against unseen attacks. Even though these methods performed better than binary classifiers in an unseen attack scenario, the performance in known attack protocols was inferior to that of binary classifiers. Xiong *et al.* [62] proposed unseen PAD methods using auto-encoders and one class classifiers with texture features extracted from images. However, the performance of the methods compared to recent CNN based methods is very poor. CNN based methods outperform most of the feature-based baselines for PAD task.

Hence there is a clear need of one class classifiers or anomaly detectors in the CNN framework. One of the drawbacks of one class model is that they do not use the information provided by the known attacks. An anomaly detector framework which utilizes the information from the known attacks could be more efficient.

Perera and Patel [50] presented an approach for one-class transfer learning in which labelled data from an unrelated task is used for feature learning. They used two loss functions, namely descriptive loss, and compactness loss to learn the representations. The data from the class of interest is used to calculate the compactness loss whereas an external multi-class dataset is used to compute the descriptive loss. Accuracy of the learned model in classification using another database is used as the descriptive loss. However, in the face PAD problem, this approach would be challenging since the *bonafide* and attack classes appear very similar.

Fatemifar *et al.* [16] proposed an approach to ensemble multiple one-class classifiers for improving the generalization of PAD. They introduced a class-specific normalization scheme for the one class scores before fusion. Seven regions, three one class classifiers and representations from three CNNs were used in the pool of classifiers. Though their method achieved better performance as compared to client independent thresholds, the performance is inferior to CNN based state of the art methods. Specifically, many CNN based approaches have achieved 0% HTER in Replay-Attack and Replay-Mobile datasets. Moreover, the challenging unseen attack scenario is not evaluated in this work.

Pérez-Cabo *et al.* [51] proposed a PAD formulation from an anomaly detection perspective. A deep metric learning model is proposed, where a triplet focal loss is used as a regularization for 'metric-softmax', which forces the network to learn discriminative features. The features learned in such a way is used together with an SVM with RBF kernel for classification. They have performed several experiments on an aggregated RGB only datasets showing the improvement made by their proposed approach. However, the analysis is mostly limited to RGB only models and 2D attacks. Challenging 3D and partial attacks are not considered in this work. Specifically, the effectiveness in challenging unknown attacks (2D vs 3D) is not evaluated.

Recently, Liu *et al.* [37] proposed an approach for the detection of unknown spoof attacks as Zero-Shot Face Anti-spoofing (ZSFA). They proposed a Deep Tree Network (DTN) which partitions the attack samples into semantic subgroups in an unsupervised manner. Each tree node in their network consists of a Convolutional Residual Unit (CRU) and a Tree Routing Unit (TRU). The objective is to route the unknown attacks to the most proper leaf node for correctly classifying it. They have considered a wide variety of attacks in their approach and their approach achieved superior performance compared to the considered baselines.

Jaiswal *et al.* [30] proposed an end to end deep learning model for PAD which used unsupervised adversarial invariance. In their method, the discrim-

inative information and nuisance factors are disentangled in an adversarial setting. They showed that by retaining only discriminative information, the PAD performance improved for the same base architecture. Mehta *et al.* [43] trained an Alexnet model with a combination of cross-entropy and focal losses. They extracted the features from Alexnet and trained a two-class SVM for PAD task. However, results in challenging datasets such as OULU and SiW were not reported.

Recently Joshua and Jain [14] utilized multiple GANs for spoof detection in fingerprints. Their method essentially consisted of training a DCGAN [53] using only the *bonafide* samples. At the end of the training, the generator is discarded, and the discriminator is used as the PAD classifier. They combined the results from different GANs operating on different features. However, this approach may not work well for face images as the recaptured images look very similar to the *bonafide* samples.

2.4 Multi-channel based approaches for face PAD

In general, most of the visible spectrum based PAD methods try to detect the subtle differences in image quality when it is recaptured. However, this method could fail as the quality of capturing devices and printers improves. For 3D attacks, the problem is even more severe. As the technology to make detailed masks is available, it becomes very hard to distinguish between *bonafide* and presentation attacks by just using visible spectrum imaging. Many researchers have suggested using multi-spectral and extended range imaging to solve this issue [54], [57].

Raghavendra *et al.* [54] presented an approach using multiple spectral bands for face PAD. The main idea is to use complementary information from different bands. To combine multiple bands they observed a wavelet-based feature level fusion, and a score fusion methodology. They experimented with detecting print attacks prepared using different kinds of printers. They obtained better performance with score level fusion as compared to the feature fusion strategy.

Erdogmus and Marcel [15] evaluated the performance of a number of face PAD approaches against 3D masks using 3DMAD dataset. This work demonstrated that 3D masks could fool PAD systems easily. They achieved HTER of 0.95% and 1.27% using simple LBP features extracted from color and depth images captured with Kinect.

Steiner *et al.* [57] presented an approach using multi-spectral SWIR imaging for face PAD. They considered four wavelengths - $935nm$, $1060nm$, $1300nm$ and $1550nm$. In their approach, they trained a SVM for classifying each pixel as a skin pixel or not. They defined a Region Of Interest (ROI) where the skin is likely to be present, and skin classification results in the

ROI is used for classifying PAs. The approach obtained 99.28 % accuracy in per pixel skin classification.

Dhamecha *et al.* [13] proposed an approach for PAD by combining the visible and thermal image patches for spoofing detection. They classified each patch as either *bonafide* or attack and used the *bonafide* patches for subsequent face recognition pipeline.

In [6] Bhattacharjee *et al.* showed that it is possible to spoof commercial face recognition systems with custom silicone masks. They also proposed to use mean temperature of face region for PAD.

Bhattacharjee *et al.* [5] presented a preliminary study of using multi-channel information for PAD. In addition to visible spectrum images, they considered thermal, near infrared, and depth channels. They showed that detecting rigid masks and 2D attacks is simple in thermal and depth channels respectively. Most of the attacks can be detected with a similar approach with combinations of different channels, where the features and combinations of channels to use are found using a learning-based approach.

Wang *et al.* [58] proposed multimodal face presentation attack detection with a ResNet based network using both spatial and channel attentions. Specifically, the approach was tailored for the *CASIA-SURF* [67] database which contained RGB, near-infrared and depth channels. The proposed model is a multi-branch model where the individual channels and fused data are used as inputs. Each input channel has its own feature extraction module and the features extracted are concatenated in a late fusion strategy. Followed by more layers to learn a discriminative representation for PAD. The network training is supervised by both center loss and softmax loss. One key point is the use of spatial and channel attention to fully utilize complementary information from different channels. Though the proposed approach achieved good results in the *CASIA-SURF* database, the challenging problem of unseen attack detection is not addressed.

Parkin *et al.* [47] proposed a multi-channel face PAD network based on ResNet. Essentially, their method consists of different ResNet blocks for each channel followed by fusion. Squeeze and excitation modules (SE) are used before fusing the channels, followed by remaining residual blocks. Further, they add aggregation blocks at multiple levels to leverage inter-channel correlations. Their approach achieved state of the art results in *CASIA-SURF* [67] database. However, the final model presented in is a combination of 24 neural networks trained with different attack specific folds, pre-trained models and random seeds, which would increase the computation greatly.

2.5 Challenges in PAD

In general, presentation attack detection in real-world scenario is challenging. Most of the PAD methods available in prevailing literature try to solve the

problem for a limited number of presentation attack instruments. Though some success has been achieved in addressing 2D presentation attacks, the performance of the algorithms in realistic 3D masks and other kinds of attacks is poor.

As the quality of attack instruments evolves, it becomes increasingly difficult to discriminate between *bonafide* and PAs in the visible spectrum alone. In addition, more sophisticated attacks, like 3D silicone masks, make PAD in visual spectra challenging. These issues motivate the use of multiple channels, making PAD systems harder to by-pass.

We argue that the accuracy of the PAD methods can get better with a multi-channel acquisition system. Multi-channel acquisition from consumer-grade devices can improve the performance significantly. Hybrid methods, combining both extended hardware and software could help in achieving good PAD performance in real-world scenarios. We extend the idea of a hybrid PAD framework and develop a multi-channel framework for presentation attack detection. Even with multi-channel methods, to achieve robustness against unseen attacks, the classifier part should move away from the typical binary classification formulation. One class classifiers could be a good alternative for binary classification in the PAD task. However, the features used for one class classifiers should be discriminative and compact to outperform binary classification.

3 Proposed method

A Multi-Channel Convolutional Neural Network (MC-CNN) based approach using a new loss function is proposed for PAD. Different stages of the framework are described below.

3.1 Preprocessing

Face detection is performed in the color channel using the MTCNN algorithm [66]. Once the face bounding box is obtained, face landmark detection is performed in the detected face bounding box using Supervised Descent Method (SDM) [63]. Alignment is accomplished by transforming image, such that the eye centers and mouth center are aligned to predefined coordinates. The aligned face images are converted to grayscale, and resized, to the resolution of 128×128 pixels. An example of the result of this first stage in the preprocessing pipeline is shown in Figure 2.

The preprocessing stage for non-RGB channels requires the images from different channels to be aligned both spatially and temporally with the color channel. For these channels, the facial landmarks detected in the color channel

are reused, and a similar alignment procedure is performed. A normalization using Mean Absolute Deviation (MAD) [33] is performed to cast the type of non-RGB facial images to 8-bit format.

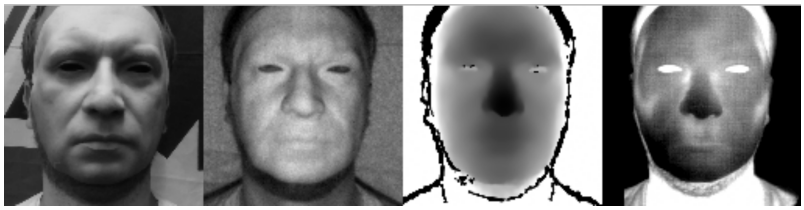


Fig. 2 Preprocessed images from a rigid mask attack; channels showed are gray-scale, infrared, depth, and thermal, respectively. Channels were preprocessed with face detection, alignment and normalization.

3.2 Network architecture

Many of previous work in face presentation attack detection utilize transfer learning from pretrained face recognition networks. This is required since the data available for PAD task is often of a very limited size, being insufficient to train a deep architecture from scratch.

The features learned in the low level of CNN networks are usually similar to Gabor filter masks, edges and blobs [65]. Deep CNNs compute more discriminant features as the depth increases [40]. It has been observed in different studies [65] and [35], that features, which are closer to the input are more general, while features in the higher levels contain task specific information. Hence, most of the literature in the transfer learning attempts to adapt the higher level features for the new tasks.

Recently, Freitas Pereira *et al.* [17] showed that the high level features in deep convolutional neural networks, trained in visual spectra, are domain independent, and they can be used to encode face images collected from different image sensing domains. Their idea was to use the shared high level features for heterogeneous face recognition task, retraining only the lower layers. In their method they split the parameters of the CNN architecture into two, the higher level features are shared among the different channels, and the lower level features (known as Domain Specific Units (DSU)) are adapted separately for different modalities. The objective was to learn the same face encoding for different channels, by adapting just the DSUs. The network was trained using contrastive loss (with Siamese architecture) or triplet loss. Retraining of only low level features has the advantage of modifying a minimal set of parameters.

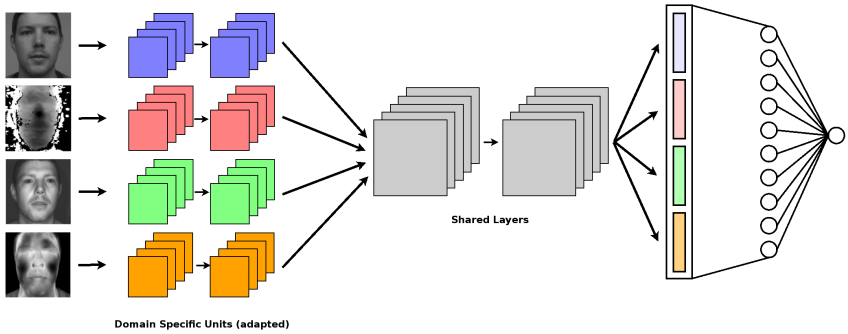


Fig. 3 Block diagram of the basic multichannel network. The gray color blocks in the CNN part represent layers which are not retrained, and other colored blocks represent re-trained/adapted layers.

We extend the idea of domain specific units (DSU) for multi-channel PAD task. Instead of forcing the representation from different channels to be same, we leverage the complementary information from a joint representation obtained from multiple channels. We hypothesize, that the joint representation contains discriminatory information for PAD task. By concatenating the representation from different channels, and using fully connected layers, a decision boundary for the appearance of *bonafide* and attack presentations can be learned via back-propagation. The lower layer features, as well as the higher level fully connected layers, are adapted in the training.

In this work, we utilize a LightCNN model [61], which was pre-trained on a large number of face images for face recognition. The LightCNN network is especially interesting as the number of parameters is much smaller than in other networks used for face recognition. LightCNN achieves a reduced set of parameters using a Max-Feature Map (MFM) operation as an alternative to Rectified Linear Units (ReLU), which suppresses low activation neurons in each layer.

The block diagram of the proposed framework is shown in Fig. 3. The pre-trained LightCNN model produces a 256-dimensional embedding, which can be used as face representation. The LightCNN model is extended to accept four channels. The 256-dimensional representation from all channels are concatenated, and two fully connected layers are added at the end for PAD task. The first fully connected layer has ten nodes, and the second one has one node. A sigmoidal activation function is used in each fully connected layer. The higher level features are more related to the task to be solved. Hence, the fully connected layers added on top of the concatenated representations are tuned exclusively for PAD task. Reusing the weights from a network pre-trained for face recognition on a large set of data, we avoid plausible over-fitting, which can occur due to limited amount of training data.

Binary Cross Entropy (BCE) is used as the loss function to train the model using the ground truth information for PAD task.

Several experiments were done by adapting the different blocks of layers, starting from the low-level features. The final fully connected layers are adapted for PAD task in all the experiments.

While doing the adaptation, the weights are always initialized from the weights of the pre-trained layers. Apart from the layers adapted, the parameters for the rest of the network remain shared.

The layers corresponding to the color channel are not adapted since the representation from the color channel can be reused for face recognition, hence making the framework suitable for simultaneous face recognition and presentation attack detection.

3.3 One Class Contrastive Loss (OCCL)

From a practical viewpoint, it is not possible to anticipate all the possible types of attacks and to have them in the training set. This, in turn, make the PAD task an unseen classification problem in a broad sense. In general, we can even consider attacks coming from different replay devices as unseen attacks. Typically, one class classifiers are well suited for such outlier detection tasks. However, in practice, the performance of one class classifiers are inferior compared to binary classifiers for known attacks, since they do not leverage useful information from the known attacks. Ideally, the PAD system should perform well in both known and unseen attack scenarios.

Clearly, there is a necessity of a method which can learn a compact one class representation while utilizing the discriminative information from known attacks. While the collection of attacks could be difficult and costly, collecting *bonafide* samples are rather easy. A new classification strategy is required to handle the realistic scenario where a limited variety of attack classes are available.

Though one class classifiers (OCC) offers a way to model the *bonafide* class, the efficient use of OCC requires the feature representation to be compact while containing discriminative information for PAD task. In the proposed framework, we use a CNN based approach to learn the feature representation. A novel loss function is proposed to learn a representation of *bonafide* samples leveraging the known attack classes.

Consider a typical CNN architecture for PAD, where the output layer contains one node and the loss function used is Binary Cross Entropy (BCE), which is defined as:

$$\mathcal{L}_{BCE} = -(y \log(p) + (1 - y) \log(1 - p)) \tag{1}$$

where y is the ground truth, ($y = 0$ for attack and $y = 1$ for *bonafide*) and p is the probability.

When trained only with BCE loss, the network learns a decision boundary based on the *bonafide* and attacks present in the training set. However, it may

not generalize when encountered with an unseen attack in the test time as it could be over-fitted to attacks which are 'known' from the training set.

To overcome this issue, we propose the 'One-Class Contrastive Loss' ($OCCL$) function which operates on the embedding layer. Proposed One-Class Contrastive Loss ($OCCL$) function is used as an auxiliary loss function in conjunction with binary cross-entropy loss. The feature map obtained from the penultimate layer of the CNN is used as the embedding. The loss function is inspired from center-loss [60] and contrastive loss [24], which are usually used in the face recognition applications.

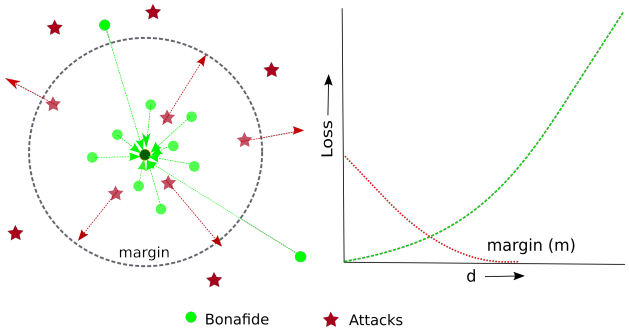


Fig. 4 Loss functions acting on the embedding space, left) *bonafide* representations are pulled closer to the center of *bonafide* class (green), while the attack embeddings(red) are forced to be beyond the margin. The attack samples outside the margin does not contribute to the loss, right) The loss as a function of distance from the *bonafide* center.

In face recognition applications, center loss is used as an additional auxiliary loss function, the task of the center loss is to minimize the distance of the embeddings from their corresponding class centers. The center loss is defined as:

$$\mathcal{L}_{center} = \frac{1}{2}\sum_{i=1}^{m} \|x_i - c_{y_i}\|_2^2 \tag{2}$$

Where L_{center} denotes the center loss, m the number of training samples in a mini-batch, $x_i \in R_d$ denotes the i^{th} training sample, y_i denotes the label, and c_{y_i} denotes the y_i^{th} class center in the embedding space.

The main issue with center loss in the PAD application is that the loss function penalizes for large intra-class distances and does not care about the inter-class distances. Contrastive center loss [52] tries to solve this issue by adding the distance between classes (inter-class) in the formulation. However, for the PAD problem, modeling the attack class as a cluster and finding a center for the attack class is not trivial. The attacks could be of different categories: 2D, 3D, and partial attacks, and it is not ideal forcing them to cluster together in the embedding space. It is only necessary to have the

embeddings of attacks far from *bonafide* cluster in the embedding space. Hence, we put the compactness constraint only on the *bonafide* class, while forcing the embeddings of PAs to be far from that of *bonafide*.

To formulate the loss function, we start with the equation for contrastive loss function proposed by Lecun *et al.* [24].

$$\begin{aligned}\mathcal{L}_{Contrastive}(W, Y, X^1, X^2) =& (1-Y)\frac{1}{2}D_W^2 \\ &+ Y\frac{1}{2}max(0, m - D_W)^2\end{aligned} \tag{3}$$

Where W is the network weights, X^1, X^2 are the pairs and Y the label of the pair, i.e., whether they belong to the same class or not. m is the margin, and D_W is the distance function between two samples. The data is provided as pairs (X^1, X^2) and the distance function D_W can be computed as the Euclidean distance.

$$D_W = \sqrt{\|X^1 - X^2\|_2^2} \tag{4}$$

Now, in our loss formulation, the critical difference is how we define D_W. In the original contrastive loss, D_W is the distance between samples. In our case, we need the representation of *bonafide* samples to be compact in an embedding space. At the same time, we want to maximize the distance between *bonafide* cluster and attack samples in the embedding space. This can be achieved by defining DC_W to be the distance from the center of *bonafide* class as follows.

$$DC_W = \sqrt{\|X^i - c_{BF}\|_2^2} \tag{5}$$

Where X^i is the embedding for i^{th} sample, and c_{BF} is the center of *bonafide* class in the embedding space.

The center of the *bonafide* class is updated in every mini-batch during training as follows.

$$c_{BF} = \hat{c}_{BF}(1-\alpha) + \alpha\frac{1}{N}\sum_{i=1}^{N} e_i \tag{6}$$

Where c_{BF} and $\hat{c}_{BF}$ denotes the new and old *bonafide*-centers. α is a scalar which prevents sudden changes in the class centers in mini-batch. e_i denotes the difference between embeddings for the *bonafide* samples in the current mini-batch compared to the previous center, and N denotes the number of *bonafide* samples in the mini-batch.

Combining the equations, our auxiliary loss function becomes:

$$\begin{aligned}\mathcal{L}_{OCCL}(W, Y, X) =& Y\frac{1}{2}DC_W^2 \\ &+ (1-Y)\frac{1}{2}max(0, m - DC_W)^2\end{aligned} \tag{7}$$

Where DC_W denotes the Euclidean distance between the samples and the *bonafide* class center, Y denotes the ground truth, i.e., $Y = 0$ for attacks and $Y = 1$ for *bonafide* (note the change in labels from the standard notation due to the ground truth convention). It is to be noted that, the proposed loss function ***does not*** require pairs of samples, which is a requirement in usage of contrastive loss. This makes it easier to train the model without requiring an explicit selection of pairs during training.

This auxiliary loss makes the representation of *bonafide* compact pushing it closer to the center of *bonafide* class and penalizes attack samples which are closer than the margin m. Attack samples which are farther than the margin m are not penalized. An illustration of the loss functions acting on the embeddings of *bonafide* and attack samples are shown in Fig. 4.

We combine the proposed loss function with standard binary cross entropy for training. The combined loss function to minimize is given as:

$$\mathcal{L} = (1-\lambda)\mathcal{L}_{BCE} + \lambda\mathcal{L}_{OCCL} \tag{8}$$

Where $\mathcal{L}$ denotes the total loss for the CNN. $\mathcal{L}_{BCE}$ and $\mathcal{L}_{OCCL}$ denotes the binary cross entropy, and one-class contrastive loss respectively. λ denotes a scalar value to set the weight for each loss functions. In our experiments we set the value of λ as 0.5.

The combined loss function $\mathcal{L}$ tries to learn a decision boundary between the available attacks and *bonafide* while the auxiliary loss tries to make the feature representation of the *bonafide* compact in the embedding space. We expect the decision boundary learned in this fashion to be more robust in unseen attacks compared to the network learned only with *BCE*. The embedding obtained in this manner is used with a one-class classifier for the PAD task.

An illustration of the proposed framework is shown in Fig. 5. At the time of training, both losses are used, and the model corresponding to the lowest validation score is selected. It is to be noted that, at the time of CNN training, both *bonafide* and (known) attack samples are used. After the CNN training, the network weights are frozen, and the *bonafide* samples are feed-forwarded to obtain the embeddings.

3.3.1 One-Class Gaussian Mixture Model

After the training of *MCCNN* with *BCE* and *OCCL*, the trained weights of the network are frozen, and it is used as a fixed feature extractor for the PAD task. Now that a compact representation is available, the objective is

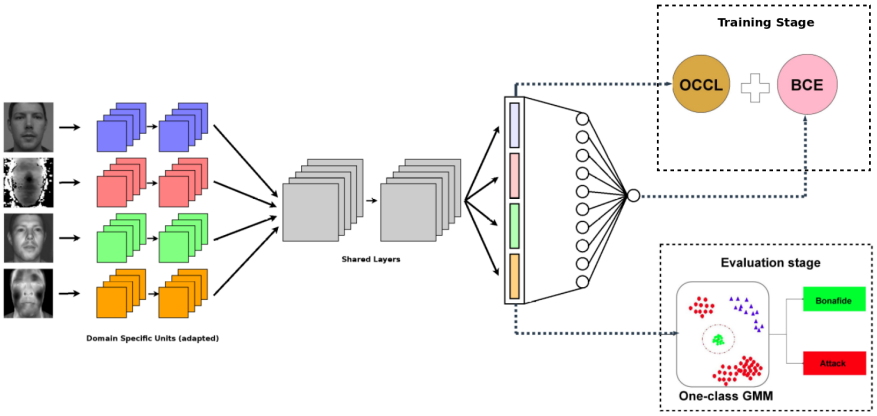


Fig. 5 Schematic diagram of the proposed framework. The CNN architecture is trained with two losses and then used as a fixed feature extractor with frozen weights. The one-class GMM is trained using the embeddings obtained from *bonafide* class alone.

to learn a one-class classifier using the features obtained. We use One-Class Gaussian Mixture Model for this task. The one class GMM is a generative approach which is used for modeling the distribution of the *bonafide* class in the proposed framework.

A Gaussian Mixture Model is defined as the weighted sum of K multivariate Gaussian distributions as:

$$p(x|\Theta) = \sum_{k=1}^{K} w_k \mathcal{N}(x; \mu_k, \Sigma_k), \tag{9}$$

where $\Theta = \{w_k, \mu_k, \sigma_k\}_{\{k=1,\ldots,K\}}$ are the weights, means and the covariance matrix of the GMM.

Expectation-Maximization (EM) [12] was used to compute the parameters of the GMM. A full covariance matrix is computed for each component, and the number of components to use was empirically selected as five ($K = 5$).

During the training phase, embeddings obtained from *bonafide* class only are used to train the One-Class GMM.

In test time, a sample is first forwarded though the network to obtain the embedding x, and then fed to the One-Class GMM to obtain the log-likelihood score as follows:

$$score = log(p(x|\Theta)) \tag{10}$$

In summary, the proposed framework can be considered as a one-class classifier based framework for PAD. The crucial distinction is that, the features used are ***learned***. The loss function proposed forces the CNN to learn a compact representation for the *bonafide* class leveraging the information from

Algorithm 1: Algorithm for training the proposed framework

Data: (x_i, y_i), where x_i is multi-channel input and $y_i \in 0, 1$; 0 – for attack and 1– for *bonafide*
Result: W_C – CNN weights, Θ_{GMM} – Parameters of GMM
Constants : λ – weighting factor, μ – learning rate
Initialize : C_{BF} – center of *bonafide* class, W_C – initial weights of CNN from pretrained model
for *mini-batch* $\leftarrow$ *1* ***to*** *P* **do**
    Forward x_i through the CNN
    Compute the combined loss: $\mathcal{L} = (1-\lambda)\mathcal{L}_{BCE} + \lambda\mathcal{L}_{OCCL}$
    Back-propagate the loss and update the weights of DSUs and FC layers
    Update the *bonafide* center:
    $c_{BF} = \hat{c}_{BF}(1-\alpha) + \alpha\frac{1}{N}\sum_{i=1}^{N} e_i$
end
Forward x_j (*bonafide*, where $y_j = 1$) through the CNN to obtain Embeddings E_j
Estimate parameters of GMM from E_j:
$\Theta_{GMM} = (w_k, \mu_k, \Sigma_k)$
Parameters$\leftarrow (W_C, \Theta_{GMM})$

known attack classes. The algorithm for training the framework is shown in Algorithm 1.

3.4 Implementation details

To increase the number of samples, data augmentation using random horizontal flips with a probability of 0.5 was used in training. Adam Optimizer [31] was used to minimize the combined loss function. Learning rate of 1×10^{-4} and a weight decay parameter of 1×10^{-5} was used. The network was trained for 50 epochs on GPU grid with a batch size of 32. The model corresponding to minimum validation loss in the *dev* set is selected as the best model. For the four-channel models, the MCCNN architecture has about 13.1M parameters and about 14.5 GFLOPS. The implementation was done using PyTorch [48] library.

4 Experiments

In order to evaluate the effectiveness of the proposed approach, we have performed experiments in three publicly available databases, namely *WMCA* [23], *MLFP* [1], and *SiW-M* [37] datasets. Recently published *CASIA-SURF* [67] database also consists of multi-channel data, namely color, depth, and infrared channels with a limited set of attack instruments. However, the raw

data from the sensors were not publicly available; in the publicly available version of the database, images were masked and scaled with custom preprocessing reducing the dynamic range of depth and infrared channels severely. Moreover, there was no guaranteed alignment between the channels. Therefore we can't use our framework with *CASIA-SURF* database due to the mentioned limitations.

4.1 WMCA dataset

Fig. 6 Attack categories in *WMCA* dataset, only RGB images are shown. Print and Replay constitutes the 2D attacks and all others are 3D attacks (Image taken from [23]).

We have conducted an extensive set of experiments on *Wide Multi-Channel presentation Attack* (*WMCA*) [3] database, which contains a total of *1679* video samples of *bonafide* and attack attempts from *72* identities. The database contains information from four different channels collected simultaneously, namely, color, depth, infrared, and thermal channels. The data was collected using two consumer devices, Intel® RealSense™SR300 capturing RGB-NIR-Depth streams, and Seek Thermal CompactPRO for the thermal channel. The database contained around eighty different PAIs constituting seven different categories of attacks: print, replay, funny eyeglasses, fake head, rigid mask, flexible silicone mask, and paper masks. The RGB visualization of the attack categories is shown in Fig. 6 and the different sessions in Fig. 7.

[3] Database available at : `https://www.idiap.ch/dataset/wmca`

Fig. 7 Different sessions in *WMCA* dataset, only RGB images are shown. A total of six sessions was used the *WMCA* (Image taken from [23])

Detailed information about the *WMCA* database can be found in the publication [23]. The statistics of the number of samples in each category and their types are shown in Table 1. We have made challenging protocols in the *WMCA* dataset to perform an extensive set of evaluations emulating real-world unseen attack scenarios.

4.1.1 Protocols in *WMCA*

To test the performance of the algorithm in known and unseen attack scenarios, we created three protocols in the *WMCA* dataset. The protocols are described below.

- **grandtest** : This is the exact same *grandtest* protocol available with *WMCA* database, here all the attack types are present in almost equal proportions in the *train*, *development* and *evaluation* sets. The attack types and *bonafide* samples are divided into three folds, and the client ids are disjoint across the three sets. Each presentation attack instrument had a separate client id. The train, dev, eval splits were made in such a way that a specific PA instrument will appear in only one fold.
- **unseen-2D** : In this protocol, we use same splits as *grandtest* and removed all 2D attacks from *train* and *development* groups. *Evaluation* set

contains only *bonafide* and 2D attacks. This emulates the performance of a system when encountered with 2D attacks which was not seen in training.

- **unseen-3D** : In this protocol, we use same splits as *grandtest* and removed all 3D attacks from *train* and *development* groups. *Evaluation* set contains only *bonafide* and 3D attacks. This emulates the performance of a system when encountered with 3D attacks which were not seen in training. This is the most challenging protocol as the model sees only the simpler 2D attacks in training and encounter challenging 3D attacks in testing.

While the *grandtest* protocol emulates the known attack scenario, other protocols emulate the unseen attack scenario. All protocols are made available publicly.

Table 1 Statistics of attacks in *WMCA* database

PA Category	Type	#Presentations
bonafide	-	347
glasses	Partial	75
print	2D	200
replay	2D	348
fake head	3D	122
rigid mask	3D	137
flexible mask	3D	379
paper mask	3D	71
TOTAL		**1679**

4.2 MLFP dataset

MLFP dataset [1] consists of attacks captured with seven 3D latex masks and three 2D print attacks. The dataset contains videos captured from color, thermal and infrared channels. Since channels were captured individually in different recording sessions, multi-channel approaches are not trivial. Also, the alignment of channels is not possible since they are not collected simultaneously. Hence, we only use the RGB videos from the MLFP dataset for our experiments. The database contains videos of 10 subjects wearing both print and latex masks. There are 440 videos are consisting of both attacks and *bonafide* for the RGB channel.

4.2.1 Protocols in *MLFP*

To emulate known and unseen attack scenarios, we created three new protocols in the *MLFP* dataset. There are two types of attacks, namely print and mask. Only two sets, i.e., *train* and *evaluation* are created due to the small size of the dataset. We used a subset of the train set (10%) for model selection. The protocols are described below.

- **grandtest** : This protocol emulates the known attack scenario. Both the attacks are present in both *train* and *evaluation* set. However, the subjects and the PAs are disjoint across the two sets.
- **unseen-print** : In this protocol, only *bonafide* and mask attacks are present in *train* set; the *evaluation* set contains only *bonafide* and print attacks. This emulates unseen attack scenario.
- **unseen-mask** : In this protocol, only *bonafide* and print attacks are present in *train* set; the *evaluation* set contains only *bonafide* and mask attacks. This protocol also emulates unseen attack scenario.

4.3 SiW-M dataset

The Spoof in the Wild database with Multiple Attack Types (*SiW-M*) [37] consists of a wide variety of attacks captured only in RGB spectra. The database consists of images from 493 subjects, and a total of 660 *bonafide* and 968 attack samples. A total of 1628 files, consisting of 13 different attack types, collected in different sessions, pose, lighting, and expression (PIE) variations. The attacks consist of various types of masks, makeups, partial attacks, and 2D attacks. The videos are available in 1080P resolution.

4.3.1 Protocols in *SiW-M*

To emulate unseen attack scenarios, we use the leave-one-out (LOO) testing protocols available with the *SiW-M* [37] dataset. The protocols consists of only *train* and *eval* sets. In each LOO protocol, the training set consists of 80% percentage of the live data and 12 types of spoof attacks. The evaluation set consists of 20% of *bonafide* data and the attack which was left out in the training phase. The subjects in *bonafide* sets are disjoint in *train* and *evaluation* sets. A subset of the train set (5%) was used for model selection. Additionally, we have created a *grandtest* protocol, specifically for cross-database testing which contains all the attack types in all the folds.

4.4 Evaluation metrics

We report the standardized ISO/IEC 30107-3 metrics [29], Attack Presentation Classification Error Rate (APCER), and Bonafide Presentation Classification Error Rate (BPCER), and Average Classification Error Rate (ACER) in the *test* set. A BPCER threshold of 1% is used for computing the threshold in *dev* set. The APCER and BPCER in both *dev* and *eval* sets are also reported. Additionally, the ROC curves for experiments are also shown in all the protocols. For the *MLFP* dataset, we report only EER in the *evaluation* set since only two sets are available. For SiW-M database, we apply a threshold selected a-priori in all protocols, for computing the metrics, to be comparable with the results in [37].

4.5 Baselines

We have implemented three feature-based baselines and two CNN based baselines. For a fair comparison, all the benchmarks are multi-channel methods and use the same four channels. Besides, an RGB only CNN model is also added for comparison. A short description of the baselines along with the acronyms used are shown below:

- *MC-RDWT-Haralick-SVM*: This baseline is the multi-channel extension of the RDWT-Haralick-SVM approach proposed in [1]; the images from all channels are stacked together after preprocessing. For each channel, the image is divided into a 4×4 grid, and Haralick [25] features obtained from the RDWT decompositions are concatenated from all the grids in all channels to get the joint feature vector. The joint feature is used with a linear SVM for PAD.
- *MC-RDWT-Haralick-GMM*: Here, the feature extraction stage is same as *MC-RDWT-Haralick-SVM*; however, the classifier used is one class GMM. Only *bonafide* samples are used in training this model. This model is added to show the performance of one class models in unseen attack scenarios.
- *MC-LBP-SVM*: Here, again, the same preprocessing is performed on all the channels first. After this, Spatially enhanced histograms of LBP representation from all the component channels are computed and concatenated to a feature vector. The features extracted are fed to an SVM for PAD task.
- *DeepPixBiS* : This is a CNN based system [20] trained using both binary and pixel-wise binary loss function. This model only uses RGB information for PAD.

- *MC-ResNetPAD*: We reimplemented the architecture from [47] extending it to four channels, based on their open-source implementation [4]. This approach obtained the first place solution in the 'CASIA-SURF' challenge. For a fair comparison, instead of using an ensemble we used the best pretrained model as suggested in [47].
- *MCCNN(BCE)* : This is the multi-channel CNN system described in [23], which achieved state of the art performance in the *grandtest* protocol. The model is trained using Binary Cross-Entropy (*BCE*) loss only.

All the baseline methods described are reproducible, and the details about the parameters can be found in our open-source package [5].

4.6 Experiments and Results in *WMCA* dataset

Table 2 Performance of the baseline systems and the proposed method in *grandtest* protocol of *WMCA* dataset. The values reported are obtained with a threshold computed for BPCER 1% in *dev* set.

Method	dev (%)		test (%)		
	APCER	ACER	APCER	BPCER	ACER
MC-RDWT-Haralick-SVM	3.6	2.3	5.4	1.2	3.3
MC-LBP-SVM	3.6	2.3	8.5	0.6	4.6
MC-RDWT-Haralick-GMM	43.4	22.2	47.7	1.7	24.7
DeepPixBiS (RGB only)[20]	1.0	1.0	8.2	3.7	6
MC-ResNetPAD [47]	3.8	2.4	3.5	1.6	2.6
MCCNN(BCE)[23]	0.4	0.7	0.5	0	**0.2**
MCCNN(BCE+OCCL)-GMM	0.1	0.6	0.6	0.1	0.4

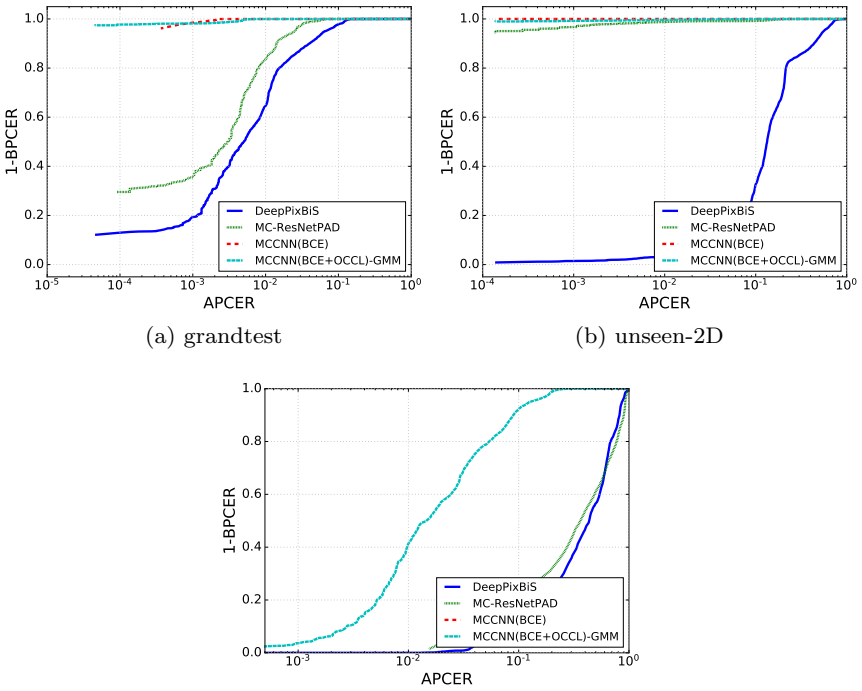
We have tested the baselines and the proposed approach in three different protocols in *WMCA*. The proposed approach is denoted as *MCCNN(BCE+OCCL)-GMM*.

- *MCCNN(BCE+OCCL)-GMM*: Here, the *bonafide* embeddings from the *MCCNN* trained using both the losses are used to train a GMM, and in the evaluation stage, the score from the one class GMM is used as the PAD score.

The results in each protocol are described below.

[4] Available from: `https://github.com/AlexanderParkin/ChaLearn_liveness_challenge`

[5] Source code: `https://gitlab.idiap.ch/bob/bob.paper.oneclass_mccnn_2019`



(a) grandtest (b) unseen-2D

(c) unseen-3D

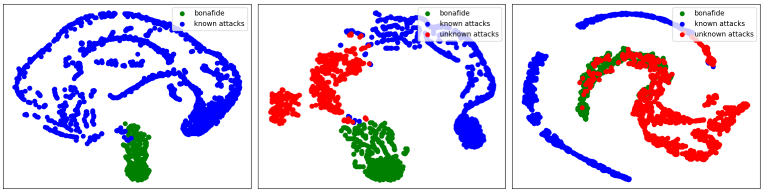
Fig. 8 DET curves for the *eval* sets of different protocols of *WMCA* dataset a) *grandtest*, b) *unseen-2D*, c) *unseen-3D* protocol.

Table 3 Performance of the baseline systems and the proposed method in **unseen** protocols of *WMCA* dataset. The values reported are obtained with a threshold computed for BPCER 1% in *dev* set.

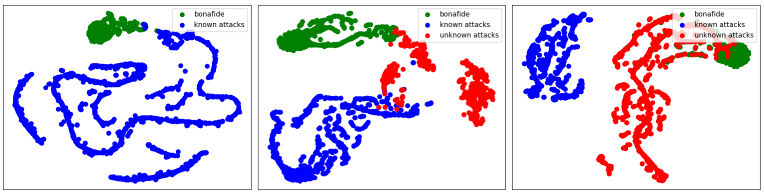
Method	unseen-2D			unseen-3D		
	APCER	BPCER	ACER	APCER	BPCER	ACER
MC-RDWT-Haralick-SVM	0.3	0.1	0.2	66.0	0.1	33.1
MC-LBP-SVM	40.7	0.1	20.4	38.9	0.2	19.5
MC-RDWT-Haralick-GMM	0.0	0.2	**0.1**	70.8	1.9	36.4
DeepPixBiS (RGB only)[20]	77.7	0.3	39	74.7	16.3	45.5
MC-ResNetPAD [47]	4.1	0.9	2.5	92.2	6.4	49.3
MCCNN(BCE)[23]	0.0	1.0	0.5	62.0	0.0	31.0
MCCNN(BCE+OCCL)-GMM	0.3	0.6	0.5	15.4	3.9	**9.7**

4.6.1 Experiments in *grandtest* protocol

The *grandtest* protocol emulates the known attack scenario. Table 2 tabulates the results in the *grandtest* protocol. The proposed approach outperforms



(a) grandtest (b) unseen-2D (c) unseen-3D



(d) grandtest (e) unseen-2D (f) unseen-3D

Fig. 9 t-SNE plots of embeddings in the protocols in *WMCA* dataset. First row (a,b,c) shows the embeddings when only *BCE* loss was used. Second row (d,e,f) shows the embeddings when both the losses are used. Embeddings of both known and unseen attacks are shown in the figures for each protocol. Grandtest protocol contains only known attacks in the test set.

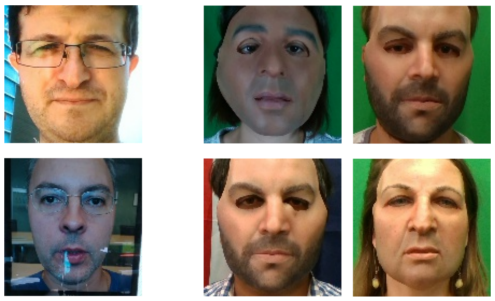


Fig. 10 The attack samples which are closer to *bonafide* cluster in a) unseen-2D (Fig. 9(E)) and b) unseen-3D ((Fig. 9(F))) protocol for the proposed framework.

the feature-based methods by a large margin as expected. The model *MC-RDWT-Haralick-GMM* trained using a one-class model achieves the worse results. It is interesting to note that the *MC-RDWT-Haralick-SVM* model, trained using the same feature as a binary classifier performed much better. This shows one weakness of one-class classifiers in a known attack scenario, as they do not use the known attacks in training. The *MCCNN(BCE)* achieves much better performance as compared to *MC-ResNetPAD*. The *MC-CNN(BCE)* trained as a binary classifier achieves the best performance in this

protocol. The proposed *MCCNN(BCE+OCCL)-GMM* approach achieves comparable performance to *MCCNN(BCE)*. This indicates that the one class GMM classifier performs on par with the binary classification, provided they are trained with compact feature representations.

4.6.2 Experiments in *unseen-2D* and *unseen-3D* protocol

The *unseen-2D* and *unseen-3D* protocols emulates the unseen attack scenario. The *unseen-3D* is the most challenging protocol since it is trained only on 2D - print and replay attacks and encounters a wide variety of 3D attacks such as silicone masks, fake heads, mannequins, etc. in the *eval* set.

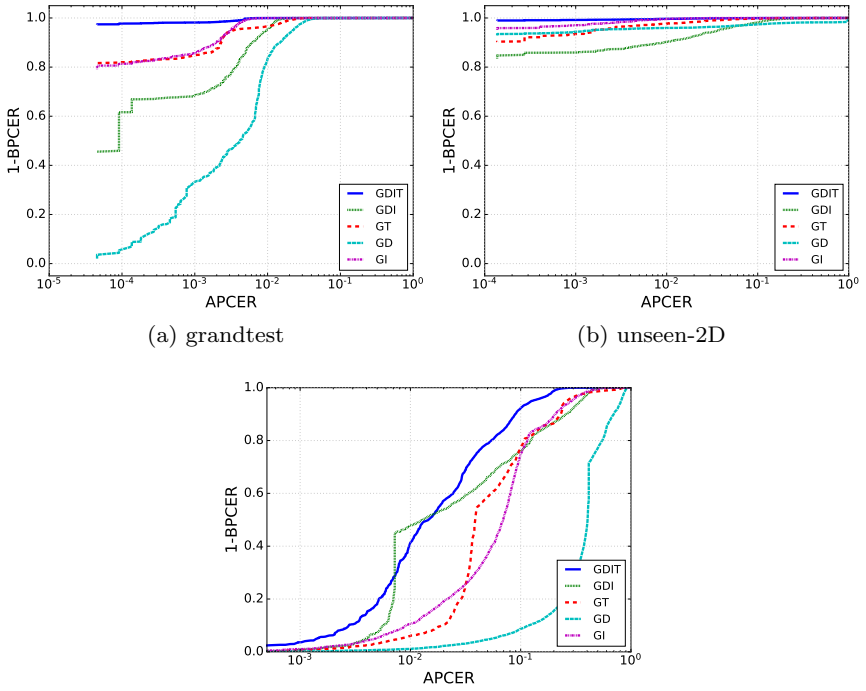
Most of the approaches perform well in the *unseen-2D* protocol. This result is intuitive as these models are trained on challenging 3D attacks, detection of 2D attacks is much easier. Moreover, the 2D attacks can be easily identified in depth, thermal, and infrared channels. Even some feature-based methods perform well in this protocol, with *MC-RDWT-Haralick-GMM* method achieving the best performance. This shows the advantage of one class model in an unseen attack scenario. The proposed approach *MCCNN(BCE+OCCL)-GMM* and *MCCNN(BCE)* baseline perform comparably in this protocol. Notably, the DeepPixBiS model achieves much worse results in this protocol. This could be because discriminating between *bonafide* and 2D attacks are harder when only RGB information is used.

The *unseen-3D* protocol shows important results. All the baselines show inferior performance when encountered with unseen 3D samples. This shows the failure of binary classifiers in generalizing to challenging unseen attacks. The *MCCNN(BCE)* approach, while being architecturally similar, fails to generalize when trained in the binary classification setting. With the proposed approach, performance improves to 9.7% when the one class GMM is used on the *bonafide* representations. Since the network learns to map the *bonafide* samples to a compact cluster in the feature space, even in the presence of unseen attacks, the decision boundary learned for the *bonafide* class is robust. The unseen attacks map far from the *bonafide* cluster and hence becomes easy to discriminate from *bonafide* samples. This result is encouraging since the network was shown only 2D attacks in training, and still, it manages to achieve good performance against challenging 3D attacks. The ROCs for all the protocols are shown in Fig. 8.

The t-SNE [38] plots of the embeddings for all protocols are shown in Fig. 9. Five frames from each video in the evaluation sets of the protocols are used for this visualization. While the difference between *bonafide* and attacks are clear in the *grandtest* and *unseen-2D*, difference in *unseen-3D* protocol is very evident. It can be clearly seen that the *bonafide* class clusters together and is far from the *bonafide* representation in the embedding space in the *unseen-3D* protocol when the proposed loss is used. Unseen attacks overlaps with *bonafide* embeddings when only *BCE* is used. This clearly demonstrates

the effectiveness of the proposed approach for unseen attack detection. The unseen attacks which are overlapping with the *bonafide* region are shown in Fig. 10. It can be seen that some video replay samples and flexible silicone 3D masks get misclassified in unseen-2D and unseen-3D protocols respectively.

4.6.3 Ablation study with channels



(a) grandtest (b) unseen-2D

(c) unseen-3D

Fig. 11 Ablation study with different combination of channels, DET curves for the *eval* sets of different protocols of *WMCA* dataset a) *grandtest*, b) *unseen-2D*, c) *unseen-3D* protocol.

To evaluate the performance of the proposed framework on different set of channels, we perform an ablation study by including a different set of channels. We used only the best performing *MCCNN(BCE+OCCL)-GMM* approach in this ablation study. In all combinations, the gray-scale channel is present since it is used as a reference. This is required as the embedding from the gray-scale part can be used for face recognition as well.

The acronyms for different channels are shown below:

- G: Gray-scale image
- D: Depth image
- I: Infrared channel
- T: Thermal channel

Various combinations of these channels are experimented with, and the results are tabulated in Table 4. It is to be noted that the channels G, D and I come from the same device and T is coming from a different device. Usually, thermal cameras are expensive, compared to RGB-D cameras, and hence the combinations involving subsets of G, D and I are more interesting from a deployment point of view.

Table 4 Performance of the proposed framework with different combinations of channels in all protocols of *WMCA* dataset. The values reported are obtained with a threshold computed for BPCER 1% in *dev* set.

Channels	grandtest	unseen-2D	unseen-3D
	ACER	ACER	ACER
GDIT	**0.4**	**0.5**	**9.7**
GDI	1.1	11.2	23.1
GT	2.2	3.2	21.5
GD	2.3	49.4	45.4
GI	1.1	2.2	22.6

From Table 4, it can be seen that the performance degrades as channels are removed. However, the combination GI achieves reasonable performance while considering the performance-cost ratio. The ROCs for different protocols are shown in Fig. 11.

4.7 Experiments and Results in *MLFP* dataset

We have used only the RGB channel for the experiments since the other channels were not captured simultaneously. For the MCCNN framework and other baselines, 'R', 'G', and 'B' are considered as the different channels in these experiments. We have performed the experiments in the three newly created protocols and the results are tabulated in Table 5.

From the results in Table 5, it can be seen that the CNN based approach outperforms the feature-based approaches. The MCCNN framework, with the addition of the newly proposed loss outperforms the architecture trained with BCE only, showing the effectiveness of the proposed approach.

Even though the proposed approach performs better than the baselines, it is to be noted that the key point of the proposed approach, leveraging multi-channel information, is not utilized here. The architecture is not optimized for PAD in RGB and this experiment is performed only to show the change in

performance with the new loss function. Nevertheless, the proposed approach achieves better performance as compared to the baselines in all the protocols.

Table 5 Performance of the proposed framework in the protocols in *MLFP* dataset. Only RGB channel was used in this experiments. The values reported are the EER in the *evaluation* set.

Algorithm	grandtest	unseen print	unseen mask
MC-RDWT-Haralick-SVM	9.8	12.0	32.2
MC-LBP-SVM	6.3	27.1	9.3
MC-RDWT-Haralick-GMM	27.4	40.8	21.5
DeepPixBiS (RGB only)[20]	6.3	24.8	17.5
MCCNN (BCE)	5.5	9.2	5.2
MCCNN (BCE+OCCL)-GMM	**1.2**	**3.3**	**3.4**

4.8 Experiments and Results in *SiW-M* dataset

Table 6 Performance of the proposed framework in the leave one out protocols in *SiW-M* dataset. Only RGB channel was present in this dataset.

Methods	Metrics	Replay	Print	Mask Attacks					Makeup Attacks			Partial Attacks			Average
				Half	Silicone	Trans.	Paper	Manne.	Obfusc.	Imperson.	Cosmetic	Funny Eye	Paper Glasses	Partial Paper	
MC-RDWT-H-SVM	ACER	17.15	16.52	22.71	22.49	24.37	8.49	10.19	42.47	8.26	25.79	33.85	24.50	9.61	20.4 ± 10.3
	EER	16.88	16.53	21.80	20.73	21.94	7.34	9.88	32.56	2.37	23.51	31.72	21.94	10.05	18.2 ± 9.0
MC-LBP-SVM	ACER	16.83	17.54	16.38	28.76	30.46	12.15	15.04	59.44	11.97	24.45	30.41	28.31	13.24	23.4 ± 12.9
	EER	15.96	16.83	16.87	28.51	29.77	10.54	12.75	52.60	1.90	24.61	28.32	26.76	11.29	21.2 ± 12.6
Auxiliary [36]	ACER	16.8	6.9	19.3	14.9	52.1	8.0	12.8	55.8	13.7	11.7	49.0	40.5	5.3	23.6 ± 18.5
	EER	14.0	4.3	11.6	12.4	24.6	7.8	10.0	72.3	10.1	9.4	21.4	18.6	4.0	17.0 ± 17.7
DTN [37]	ACER	9.8	6.0	15.0	18.7	36.0	4.5	7.7	48.1	11.4	14.2	19.3	19.8	8.5	16.8 ± 11.1
	EER	10.0	2.1	14.4	18.6	26.5	5.7	9.6	50.2	10.1	13.2	19.8	20.5	8.8	16.1 ± 12.2
DeepPixBiS [20]	ACER	13.94	8.30	6.38	16.53	35.47	4.43	4.81	55.27	5.80	18.95	37.48	43.87	3.69	19.6 ± 17.4
	EER	11.68	7.94	7.22	15.04	21.30	3.78	4.52	26.49	1.23	14.89	23.28	18.90	4.82	12.3 ± 8.2
MCCNN (BCE)	ACER	23.01	18.52	7.66	15.02	22.56	4.29	6.02	40.31	5.86	20.19	29.72	32.52	16.54	18.6 ± 11.1
	EER	17.08	11.83	7.56	12.82	16.09	0.71	6.85	25.94	2.29	16.30	18.90	22.82	13.13	13.2 ± 7.4
MCCNN(BCE+OCCL)-GMM	ACER	12.61	12.84	9.69	11.97	25.16	6.87	5.89	29.90	6.34	16.01	16.83	26.97	13.66	14.9 ± 7.8
	EER	12.82	12.94	11.33	13.70	13.47	0.56	5.60	22.17	0.59	15.14	14.40	23.93	9.82	$\mathbf{12.0 \pm 6.9}$

Table 6 shows the performance of the proposed framework, again only in the RGB scenario. CNN-based methods are much more powerful than feature-based methods in this case. It can be seen that the proposed approach achieves better performance compared to the baseline methods. The performance of the *MCCNN (BCE+OCCL)-GMM* model is better than that of the *MCCNN(BCE)* model. It can be seen that the addition of the new loss function makes the classification of unseen attacks more accurate.

4.9 Cross-database evaluations

Table 7 The results from the cross-database testing between WMCA and SiW-M datasets using the grandtest protocol, only RGB channels were used in this experiment.

Method	trained on WMCA		trained on SiW-M	
	tested on WMCA	tested on SiW-M	tested on SiW-M	tested on WMCA
MC-RDWT-Haralick-SVM	14.6	29.6	15.1	45.3
MC-LBP-SVM	26.6	45.5	19.6	38.6
MC-RDWT-Haralick-GMM	27.9	34.0	25.5	43.6
DeepPixBiS	7.5	49.1	14.7	44.4
MCCNN (BCE)	12.1	34.0	9.9	42.3
MCCNN (BCE+OCCL)-GMM	12.3	31.9	9.5	41.8

Since we cannot carry out a cross-database evaluation between a multi-channel database and an RGB-only database, we only used the RGB channels from two data sets for the cross-database evaluation. We selected WMCA and SiW-M datasets since they are relatively large and consist of a wide variety of attacks.

Table 7 shows that the MCCNN model achieves comparable performance with and without the new loss. In general, performance in the cross-database setting is poor for all models. The poor performance can be due to the disparity in acquisition conditions and attack types. A larger variety of attacks makes it more difficult for the classifier to identify attacks only via RGB channels. Cross-database performance against this multitude of attacks appears to be more challenging than typical cross-database evaluations that only use 2D attacks. Using multiple channels [23] may alleviate these issues. This also indicates the limitation of RGB-only methods while dealing with a wide variety of attacks.

5 Discussions

The experiments in the *WMCA* database clearly show that the CNN-based methods outperform the feature-based methods by a large margin. When comparing the method *MCCNN (BCE)* with the proposed method, the performance in the known attack scenario is comparable. This indicates that the proposed one-class GMM-based approach performs par with the binary classification, thanks to the embedding learned with the proposed loss function.

Most approaches work well in the *unseen-2D* protocol, as it can be clearly distinguished in many channels. Furthermore, it shows that simpler attacks are easy to spot if the network is trained in challenging attacks. While the performance in the *grandtest* and *unseen-2D* protocols is comparable, the proposed method achieves a great increase in performance in the most challenging *unseen-3D* protocol. The proposed loss function forces the network to learn a compact representation for *bonafide* examples in the feature space. Both known and unknown attacks are mapped far away from the *bonafide* cluster in the feature space. The decision boundary learned from the one-class model seems robust in identifying both seen and unseen attacks in such a scenario. This finding is significant for several reasons. It is to be noted that in the *unseen-3D* protocol, the network is trained with only 2D attacks, i.e., prints and replays. The proposed method achieves excellent performance in a test set consisting of challenging 3D attacks such as custom silicone masks, paper masks, mannequins, etc. The real-world implications of this approach are very promising. The proposed method can be used to develop robust PAD systems without the requirement of having to manufacture costly presentation attacks. Depending on availability, the PAD models can be trained using easily available attacks. The proposed framework utilizes the available (known) attack categories to learn a robust representation to facilitate known and unseen attack detection. It is to be noted that the compact representation is made possible by the joint multi-channel representation used.

In practical deployment scenarios, computational or cost constraints can prevent the use of all four channels. In such a situation, models trained on available channels can be selected based on the cost-performance ratio by sub-selecting the channels. The results of the ablation study in Table 4 can be used to determine which channels should be used in such cases.

In a similar way, the experiments in *MLFP* and *SiW-M* databases also show that CNN-based methods outperform feature-based baselines. Although we did not use multichannel information in these experiments, the experimental results show the performance improvement with the new loss function. Using the proposed framework together with network backbones designed specifically for RGB PAD might improve the results.

The cross-database performance shows the limitations of the RGB channel when tested with a wide variety of attacks. The performance of the baselines, as well as the proposed approach, is poor when only using RGB data. This shows the challenging nature of RGB only PAD while considering a multitude of attacks. Using multiple channels as done with the *WMCA* dataset might improve the performance.

6 Conclusions

Detecting face presentation attacks is often considered as a binary classification task which results in over-fitting to known attacks and results in poor generalization against unseen attacks. In this chapter, we address this problem with a new multi-channel framework that uses a one-class classifier. A novel loss function is formulated which forces the network to learn a compact yet discriminative representation for the face images. Thanks to the proposed loss function, the *bonafide* samples form a compact cluster in the feature space. A decision boundary around the representation of *bonafide* class can be obtained using the one-class model. Both known and unknown attacks are mapped far away from the *bonafide* cluster in the feature space, which can be classified by the one-class model. The proposed framework offers a new way to learn a robust PAD system from *bonafide* and available (known) attack samples. The proposed system was evaluated in the challenging datasets such as *WMCA*, *MLFP*, and *SiW-M* and was observed to surpass the baselines in both known and unseen attack scenarios. The drastic improvement in the performance in the *unseen-3D* protocol in *WMCA* shows the robustness of the proposed approach against unseen attacks thanks to the multi-channel information. The proposed method also shows an improvement even when used together with RGB channels alone. The source code and protocols to reproduce the results are made available publicly to enable further extensions of the proposed framework.

Acknowledgements Part of this research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA R&D Contract No. 2017-17020200005. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

References

1. Agarwal, A., Yadav, D., Kohli, N., Singh, R., Vatsa, M., Noore, A.: Face presentation attack with latex masks in multispectral videos. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 275–283 (2017). DOI 10.1109/CVPRW.2017.40
2. Anjos, A., Marcel, S.: Counter-measures to photo attacks in face recognition: a public database and a baseline. In: Biometrics (IJCB), 2011 international joint conference on, pp. 1–7. IEEE (2011)
3. Arashloo, S.R., Kittler, J., Christmas, W.: An anomaly detection approach to face spoofing detection: A new formulation and evaluation protocol. IEEE Access **5**, 13868–13882 (2017)

4. Atoum, Y., Liu, Y., Jourabloo, A., Liu, X.: Face anti-spoofing using patch and depth-based cnns. In: Biometrics (IJCB), 2017 IEEE International Joint Conference on, pp. 319–328. IEEE (2017)
5. Bhattacharjee, S., Marcel, S.: What you can't see can help you–extended-range imaging for 3d-mask presentation attack detection. In: Proceedings of the 16th International Conference on Biometrics Special Interest Group., EPFL-CONF-231840. Gesellschaft fuer Informatik eV (GI) (2017)
6. Bhattacharjee, S., Mohammadi, A., Marcel, S.: Spoofing deep face recognition with custom silicone masks. Biometrics Theory, Applications and Systems (BTAS), 2018 IEEE 9th International Conference on (2018)
7. Boulkenafet, Z., Komulainen, J., Akhtar, Z., Benlamoudi, A., Samai, D., Bekhouche, S.E., Ouafi, A., Dornaika, F., Taleb-Ahmed, A., Qin, L., et al.: A competition on generalized software-based face presentation attack detection in mobile scenarios. IJCB **7** (2017)
8. Boulkenafet, Z., Komulainen, J., Hadid, A.: Face anti-spoofing based on color texture analysis. In: Image Processing (ICIP), 2015 IEEE International Conference on, pp. 2636–2640. IEEE (2015)
9. Chingovska, I., Anjos, A., Marcel, S.: On the effectiveness of local binary patterns in face anti-spoofing. In: Proceedings of the 11th International Conference of the Biometrics Special Interest Group, EPFL-CONF-192369 (2012)
10. Chingovska, I., Dos Anjos, A.R.: On the use of client identity information for face antispoofing. IEEE Transactions on Information Forensics and Security **10**(4), 787–796 (2015)
11. Costa-Pazo, A., Bhattacharjee, S., Vazquez-Fernandez, E., Marcel, S.: The replay-mobile face presentation-attack database. In: Biometrics Special Interest Group (BIOSIG), 2016 International Conference of the, pp. 1–7. IEEE (2016)
12. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the em algorithm. JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B **39**(1), 1–38 (1977)
13. Dhamecha, T.I., Nigam, A., Singh, R., Vatsa, M.: Disguise detection and face recognition in visible and thermal spectrums. In: Biometrics (ICB), 2013 International Conference on, pp. 1–8. IEEE (2013)
14. Engelsma, J.J., Jain, A.K.: Generalizing fingerprint spoof detector: Learning a one-class classifier. arXiv preprint arXiv:1901.03918 (2019)
15. Erdogmus, N., Marcel, S.: Spoofing face recognition with 3d masks. IEEE transactions on information forensics and security **9**(7), 1084–1097 (2014)
16. Fatemifar, S., Awais, M., Arashloo, S.R., Kittler, J.: Combining multiple one-class classifiers for anomaly based face spoofing attack detection. In: International Conference on Biometrics (ICB) (2019)
17. de Freitas Pereira, T., Anjos, A., Marcel, S.: Heterogeneous face recognition using domain specific units. IEEE Transactions on Information Forensics and Security (2018)
18. Galbally, J., Marcel, S., Fierrez, J.: Image quality assessment for fake biometric detection: Application to iris, fingerprint, and face recognition. IEEE transactions on image processing **23**(2), 710–724 (2014)
19. Gan, J., Li, S., Zhai, Y., Liu, C.: 3d convolutional neural network based on face anti-spoofing. In: Multimedia and Image Processing (ICMIP), 2017 2nd International Conference on, pp. 1–5. IEEE (2017)
20. George, A., Marcel, S.: Deep pixel-wise binary supervision for face presentation attack detection. International Conference on Biometrics (2019)
21. George, A., Marcel, S.: Can your face detector do anti-spoofing? face presentation attack detection with a multi-channel face detector. Idiap Research Report, Idiap-RR-12-2020 (2020)

22. George, A., Marcel, S.: Learning one class representations for face presentation attack detection using multi-channel convolutional neural networks. IEEE Transactions on Information Forensics and Security pp. 1–1 (2020)
23. George, A., Mostaani, Z., Geissenbuhler, D., Nikisins, O., Anjos, A., Marcel, S.: Biometric face presentation attack detection with multi-channel convolutional neural network. IEEE Transactions on Information Forensics and Security pp. 1–1 (2019). DOI 10.1109/TIFS.2019.2916652
24. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), vol. 2, pp. 1735–1742. IEEE (2006)
25. Haralick, R.M.: Statistical and structural approaches to texture. Proceedings of the IEEE **67**(5), 786–804 (1979)
26. Heusch, G., George, A., Geissbühler, D., Mostaani, Z., Marcel, S.: Deep models and shortwave infrared information to detect face presentation attacks. IEEE Transactions on Biometrics, Behavior, and Identity Science (T-BIOM) (2020)
27. Heusch, G., Marcel, S.: Pulse-based features for face presentation attack detection. Biometrics Theory, Applications and Systems (BTAS), 2018 IEEE 9th International Conference on, Special Session On Image And Video Forensics In Biometrics (IVFIB). (2018)
28. Heusch, G., Marcel, S.: Remote blood pulse analysis for face presentation attack detection. In: Handbook of Biometric Anti-Spoofing, pp. 267–289. Springer (2019)
29. ISO/IEC JTC 1/SC 37 Biometrics: Information technology –International Organization for Standardization. Iso standard, International Organization for Standardization (2016)
30. Jaiswal, A., Xia, S., Masi, I., AbdAlmageed, W.: Ropad: Robust presentation attack detection through unsupervised adversarial invariance. arXiv preprint arXiv:1903.03691 (2019)
31. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
32. Learned-Miller, E., Huang, G.B., RoyChowdhury, A., Li, H., Hua, G.: Labeled faces in the wild: A survey. In: Advances in face detection and facial image analysis, pp. 189–248. Springer (2016)
33. Leys, C., Ley, C., Klein, O., Bernard, P., Licata, L.: Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. Journal of Experimental Social Psychology **49**(4), 764–766 (2013)
34. Li, H., He, P., Wang, S., Rocha, A., Jiang, X., Kot, A.C.: Learning generalized deep feature representation for face anti-spoofing. IEEE Transactions on Information Forensics and Security **13**(10), 2639–2652 (2018)
35. Li, H., Lu, H., Lin, Z., Shen, X., Price, B.: Lcnn: Low-level feature embedded cnn for salient object detection. arXiv preprint arXiv:1508.03928 (2015)
36. Liu, Y., Jourabloo, A., Liu, X.: Learning deep models for face anti-spoofing: Binary or auxiliary supervision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 389–398 (2018)
37. Liu, Y., Stehouwer, J., Jourabloo, A., Liu, X.: Deep tree learning for zero-shot face anti-spoofing. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
38. Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(Nov), 2579–2605 (2008)
39. Määttä, J., Hadid, A., Pietikäinen, M.: Face spoofing detection from single images using micro-texture analysis. In: Biometrics (IJCB), 2011 international joint conference on, pp. 1–7. IEEE (2011)
40. Mallat, S.: Understanding deep convolutional networks. Phil. Trans. R. Soc. A **374**(2065), 20150203 (2016)

41. Marcel, S., Nixon, M., Li, S.: Handbook of biometric anti-spoofing-trusted biometrics under spoofing attacks. Advances in Computer Vision and Pattern Recognition. Springer (2014)
42. Marcel, S., Nixon, M.S., Fierrez, J., Evans, N.: Handbook of biometric anti-spoofing : Presentation attack detection. Editors: Marcel, S., Nixon, M.S., Fierrez, J., Evans, N. (Eds.); Springer International Publishing, 2018, 2nd ed.; ISBN: 978-3319926261 (2018). DOI http://dx.doi.org/10.1007/978-3-319-92627-8. URL `http://www.eurecom.fr/publication/5667`
43. Mehta, S., Uberoi, A., Agarwal, A., Vatsa, M., Singh, R.: Crafting a panoptic face presentation attack detector
44. Mostaani, Z., George, A., Heusch, G., Geissenbuhler, D., Marcel, S.: The high-quality wide multi-channel attack (hq-wmca) database. Idiap-RR Idiap-RR-22-2020, Idiap (2020)
45. Nikisins, O., George, A., Marcel, S.: Domain adaptation in multi-channel autoencoder based features for robust face anti-spoofing. In: 2019 International Conference on Biometrics (ICB), pp. 1–8. IEEE (2019)
46. Nikisins, O., Mohammadi, A., Anjos, A., Marcel, S.: On effectiveness of anomaly detection approaches against unseen presentation attacks in face anti-spoofing. In: The 11th IAPR International Conference on Biometrics (ICB 2018), EPFL-CONF-233583 (2018)
47. Parkin, A., Grinchuk, O.: Recognizing multi-modal face spoofing with face recognition networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 0–0 (2019)
48. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch. In: NIPS-W (2017)
49. Patel, K., Han, H., Jain, A.K., Ott, G.: Live face video vs. spoof face video: Use of moiré patterns to detect replay video attacks. In: Biometrics (ICB), 2015 International Conference on, pp. 98–105. IEEE (2015)
50. Perera, P., Patel, V.M.: Learning deep features for one-class classification. IEEE Transactions on Image Processing **28**(11), 5450–5463 (2019)
51. Pérez-Cabo, D., Jiménez-Cabello, D., Costa-Pazo, A., López-Sastre, R.J.: Deep anomaly detection for generalized face anti-spoofing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 0–0 (2019)
52. Qi, C., Su, F.: Contrastive-center loss for deep neural networks. In: 2017 IEEE International Conference on Image Processing (ICIP), pp. 2851–2855. IEEE (2017)
53. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 (2015)
54. Raghavendra, R., Raja, K.B., Venkatesh, S., Busch, C.: Extended multispectral face presentation attack detection: An approach based on fusing information from individual spectral bands. In: Information Fusion (Fusion), 2017 20th International Conference on, pp. 1–6. IEEE (2017)
55. Ramachandra, R., Busch, C.: Presentation attack detection methods for face recognition systems: a comprehensive survey. ACM Computing Surveys (CSUR) **50**(1), 8 (2017)
56. Shao, R., Lan, X., Yuen, P.C.: Deep convolutional dynamic texture learning with adaptive channel-discriminability for 3d mask face anti-spoofing. In: Biometrics (IJCB), 2017 IEEE International Joint Conference on, pp. 748–755. IEEE (2017)
57. Steiner, H., Kolb, A., Jung, N.: Reliable face anti-spoofing using multispectral swir imaging. In: Biometrics (ICB), 2016 International Conference on, pp. 1–8. IEEE (2016)
58. Wang, G., Lan, C., Han, H., Shan, S., Chen, X.: Multi-modal face presentation attack detection via spatial and channel attentions. In: Proceedings of the IEEE

Conference on Computer Vision and Pattern Recognition Workshops, pp. 0–0 (2019)
59. Wen, D., Han, H., Jain, A.K.: Face spoof detection with image distortion analysis. IEEE Transactions on Information Forensics and Security **10**(4), 746–761 (2015)
60. Wen, Y., Zhang, K., Li, Z., Qiao, Y.: A discriminative feature learning approach for deep face recognition. In: European conference on computer vision, pp. 499–515. Springer (2016)
61. Wu, X., He, R., Sun, Z., Tan, T.: A light cnn for deep face representation with noisy labels. IEEE Transactions on Information Forensics and Security **13**(11), 2884–2896 (2018)
62. Xiong, F., AbdAlmageed, W.: Unknown presentation attack detection with face rgb images. In: 2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS), pp. 1–9. IEEE (2018)
63. Xiong, X., De la Torre, F.: Supervised descent method and its applications to face alignment. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 532–539 (2013)
64. Yang, J., Lei, Z., Li, S.Z.: Learn convolutional neural network for face anti-spoofing. arXiv preprint arXiv:1408.5601 (2014)
65. Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How transferable are features in deep neural networks? In: Advances in neural information processing systems, pp. 3320–3328 (2014)
66. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters **23**(10), 1499–1503 (2016)
67. Zhang, S., Wang, X., Liu, A., Zhao, C., Wan, J., Escalera, S., Shi, H., Wang, Z., Li, S.Z.: Casia-surf: A dataset and benchmark for large-scale multi-modal face anti-spoofing. arXiv preprint arXiv:1812.00408 (2018)