Idiap Research Report

ROBUST FACE PRESENTATION ATTACK DETECTION WITH MULTI-CHANNEL NEURAL NETWORKS

Anjith George[a] Sébastien Marcel

Idiap-RR-03-2022

MARCH 2022

[a]Idiap Research Institute

Robust Face Presentation Attack Detection with Multi-channel Neural Networks

Anjith George and Sébastien Marcel

Abstract Vulnerability against presentation attacks remains a challenging issue limiting the reliable use of face recognition systems. Though several methods have been proposed in the literature for the detection of presentation attacks, the majority of these methods fail in generalizing to unseen attacks and environments. Since the quality of attack instruments keeps getting better, the difference between bonafide and attack samples is diminishing making it harder to distinguish them using the visible spectrum alone. In this context, multi-channel presentation attack detection methods have been proposed as a solution to secure face recognition systems. Even with multiple channels, special care needs to be taken to ensure that the model generalizes well in challenging scenarios. In this chapter, we present three different strategies to use multi-channel information for presentation attack detection. Specifically, we present different architecture choices for fusion, along with ad-hoc loss functions as opposed to standard classification objective. We conduct an extensive set of experiments in the HQ-WMCA dataset, which contains a wide variety of attacks and sensing channels together with challenging unseen attack evaluation protocols. We make the protocol, source codes, and data publicly available to enable further extensions of the work.

1 Introduction

While face recognition technology has become a ubiquitous method for biometric authentication, the vulnerability to presentation attacks (PA) (also known as "spoofing attacks") is a major concern when used in secure scenarios [14], [17]. These attacks can be either *impersonation* or *obfuscation* attacks. Impersonation attacks

Anjith George
Idiap Research Institute, Martigny, Switzerland e-mail: anjith.george@idiap.ch

Sébastien Marcel
Idiap Research Institute, Martigny, Switzerland e-mail: sebastien.marcel@idiap.ch

attempt to gain access by masquerading as someone else and obfuscation attacks attempt to evade face recognition systems. While many methods have been suggested in the literature to address this problem, most of these methods fail in generalizing to unseen attacks [19, 56]. Another challenge is poor generalization across different acquisition settings, such as sensors and lighting. In a practical scenario, it is not possible to anticipate all the types of attacks at the time of training a presentation attack detection (PAD) model. Moreover, a PAD system is expected to detect new types of sophisticated attacks. It is therefore important to have unseen attack robustness in PAD models.

The majority of the literature deals with the detection of these attacks with RGB cameras. Over the years, many feature-based methods have been proposed using color, texture, motion, liveliness cues, histogram features [11], local binary pattern [45], [12] and motion patterns [4] for performing PAD. Recently several Convolutional Neural Network (CNN) based methods have also been proposed including 3D-CNN [21], part-based models [40] and so on. Some works have shown that using auxiliary information in the form of binary or depth supervision improves performance [6, 23]. In depth supervision, the model is trained to regress the depth map of the face as an auxiliary supervision. However, most of these methods have been designed specifically for 2D attacks and the performance of these methods against challenging 3D and partial attacks is poor [43]. Moreover, these methods suffer from poor unseen attack robustness.

The performance of RGB only models deteriorates with sophisticated attacks such as 3D masks and partial attacks. Due to the limitations of visible spectrum alone, several multi-channel methods have been proposed in literature such as [59, 17, 64, 15, 3, 8, 7, 29, 30, 24, 25, 27, 22] for face PAD. Essentially, it becomes more difficult to fool a multi-channel PAD system as it captures complementary information from different channels. Deceiving different channels at the same time requires considerable effort. Multi-channel methods have proven to be effective, but this comes at the expense of customized and expensive hardware. This could make these systems difficult to deploy widely, even if they are robust. Nevertheless, well-known commercial systems like Apple's Face ID [1] demonstrate the robustness of multi-channel PAD. A variety of channels are available for PAD, e.g., RGB, depth, thermal, near-infrared (NIR) spectra [29], shortwave infrared (SWIR) spectra [30, 64], ultraviolet [63], light field imagery [58], hyper-spectral imaging [37], etc.

Even when using multiple channels, the models tend to overfit to attacks seen in the training set. While the models could perform perfectly in attacks seen in the training set, degradation in performance is often observed when confronted with unseen attacks in real-world scenarios. This is a common phenomenon with most of the machine learning algorithms, and this problem is aggravated in case of a limited amount of training data. The models, in the lack of strong priors, could overfit to the statistical biases of specific datasets it was trained on and could fail in generalizing to unseen samples. Multi-channel methods also suffer from an increased possibility of overfitting as they increase the number of parameters due to the extra channels.

In this work, we present three different strategies to fuse multi-channel information for presentation attack detection. We consider early, late, and hybrid fusion approaches and evaluate their performance in a multi-channel setting. The joint representation helps in identifying important discriminative information in detection of the attacks.

The main contributions of this work are listed below:

- We present different strategies for the fusion of multi-channel information for presentation attack detection.
- An extensive set of experiments in the HQ-WMCA database, which contains a wide variety of attacks, using both seen and unseen attack protocols.

Additionally, the source code and protocols to reproduce the results are available publicly[1].

2 Related works

Majority of the literature in face PAD is focused on the detection of 2D attacks and uses feature-based methods [11], [45], [4],[60], [32] or CNN based methods. Recently, CNN based methods have been more successful as compared to feature-based methods [42], [23], [6], [62]. These methods usually leverage the quality degradation during 'recapture' and are often useful only for the detection of attacks like 2D prints and replays. Sophisticated attacks like 3D masks [9] are more harder to detect using RGB information alone and pose serious threat to the reliability of face recognition systems [56].

2.1 RGB Only approaches (Feature based and CNNs)

2.1.1 Feature based approaches for face PAD

For PAD using visible spectrum images, several methods such as detecting motion patterns [4], color texture and histogram based methods in different color spaces, and variants of Local Binary Patterns (LBP) in grayscale [11] and color images [12], [45] have shown good performance. Image quality based features [20] is one of the successful methods available in prevailing literature. Methods identifying moiré patterns [52], and image distortion analysis [67], 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 [49].

Chingovska *et al.* [13] 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.

[1] `https://gitlab.idiap.ch/bob/bob.paper.cross_modal_focal_loss_cvpr2021`

Arashloo *et al.* [5] 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 classifiers in the unseen attack scenario. A variant of Binarized statistical image features (BSIF), 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.* [49] 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 [12], REPLAY-MOBILE [14], and MSU-MFSD [67] datasets. A good review of related works on face PAD in color channel and available databases can be found in [61].

Heusch and Marcel [31] 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 support vector machines (SVM) for PA detection. Their approach obtained better performance than state of the art methods using rPPG in four publicly available databases.

2.1.2 CNN based approaches for face PAD

Recently, several authors have reported good performance in PAD using convolutional neural networks (CNN). Gan *et al.* [21] proposed a 3DCNN-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.* [69] 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 an SVM classifier and used for the final PAD task. Boulkenafet *et al.* [10] 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.* [39] proposed a 3D CNN architecture, which utilizes both the 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 the 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.* [42] presented a novel method for PAD with auxiliary supervision. Instead of training a network end-to-end directly for the PAD task, they used the 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 the PAD task. The addition of the auxiliary task improved the generalization capability.

Atoum *et al.* [6] 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 the entire face image. The generated depth map was converted to a 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.* [62] proposed a deep convolutional network-based architecture for 3D mask PAD. They tried to capture the subtle differences in facial dynamics using CNN. Feature maps obtained from the convolutional layer of a pre-trained VGG network were used to extract features in each channel. The 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 the 3DMAD dataset.

George *et al.* [23] presented an approach for detection of presentation attacks using a training strategy leveraging both binary and pixel-wise binary loss. The method achieved superior intra as well as cross-database performance when fine-tuned from pretrained DenseNet blocks, showing the effectiveness of the proposed loss function.

In [28], George and Marcel have shown that fine-tuning vision transformer models work well in both intra as well as cross-database settings. However, the computational complexity of these models makes it harder to deploy these models in edge devices.

2.1.3 One class classifier based approaches

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, 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.*[5] and Nikisins *et al.* [49] 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.* [68] 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 for 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 that utilizes the information from the known attacks could be more efficient.

Perera and Patel [54] presented an approach for one-class transfer learning in which labeled 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.* [18] 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% Half Total Error Rate (HTER) in Replay-Attack and Replay-Mobile datasets. Moreover, the challenging unseen attack scenario is not evaluated in this work.

Pérez-Cabo *et al.* [55] 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 are used together with an SVM with Radial Basis Function (RBF) kernel for classification. They have performed several experiments on an aggregated RGB-only dataset 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.* [44] 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 sub-groups 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 them. 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.* [36] proposed an end-to-end deep learning model for PAD that used unsupervised adversarial invariance. In their method, the discriminative 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.* [46] 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 the PAD task. However, results in challenging datasets such as OULU and SiW were not reported.

Recently Joshua and Jain [16] utilized multiple Generative Adversarial Networks (GAN) for spoof detection in fingerprints. Their method essentially consisted of training a Deep Convolutional GAN (DCGAN) [57] 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.

2.2 Multi-channel methods

In general, most of the visible spectrum-based PAD methods try to detect the subtle differences in image quality when it is recaptured. With the advances in sensor and printer technology, the quality of the generated PA instruments improve over time. The high fidelity of PAs might make it difficult to recognize the subtle differences between bonafide and PAs. 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 [59], [64].

Raghavendra *et al.* [59] 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 [17] evaluated the performance of several face PAD approaches against 3D masks using the 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.* [64] presented an approach using multi-spectral SWIR imaging for face PAD. They considered four wavelengths - 935*nm*, 1060*nm*, 1300*nm* and 1550*nm*. In their approach, they trained an 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 are used for classifying PAs. The approach obtained 99.28 % accuracy in per pixel skin classification.

Dhamecha *et al.* [15] 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 [8] Bhattacharjee *et al.* showed that it is possible to spoof commercial face recognition systems with custom silicone masks. They also proposed to use the mean temperature of the face region for PAD.

Bhattacharjee *et al.* [7] 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.* [65] 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* [71] 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.* [50] 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* [71] database. However, the final model presented 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.

2.3 Open Challenges in PAD

In general, presentation attack detection in a real-world scenario is still 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 bypass.

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 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.

3 PAD Approach

We present three different strategies to fuse multi-channel information for the presentation attack detection task. Different stages of the PAD framework are described in this section.

3.1 Preprocessing

The PAD pipeline acts on the cropped facial images. For the RGB image, the preprocessing stage consists of face detection and landmark localization using the MTCNN [70] framework, followed by alignment. The detected face is aligned by making the eye centers horizontal followed by resizing them to a resolution of 224×224. For the non-RGB images, a normalization method using the median absolute deviation (MAD) [48] is used to normalize the face image to an 8-bit range. The raw images from RGB and other channels are already spatially registered so that the same transformation can be used to align the face in the non-RGB channels.

3.2 Network Architectures for Multi-channel PAD

From the prevailing literature, it has been observed that multi-channel methods are robust against a wide range of attacks [29, 30, 24, 25]. Broadly, there are four different strategies to fuse the information from multiple channels, they are 1) early fusion, meaning the channels are stacked at the input level (for example, MC-PixBiS [30]). The second strategy is late fusion, meaning the representations from different networks are combined at a later stage similar to feature fusion (for example MC-

CNN [29]). A third strategy is a hybrid approach where information from multiple levels is combined as in [50] or [26]. A fourth strategy is score level fusion where individual networks are trained separately for different channels and score level fusion is performed on the scalar scores from each channel. However, the score fusion performs poorly compared to other methods since it does not use cross-channel relations efficiently. The details of the fusion strategies used are presented in the following sub-section.

3.2.1 Late Fusion: Multi-Channel CNN (MCCNN-OCCL-GMM)

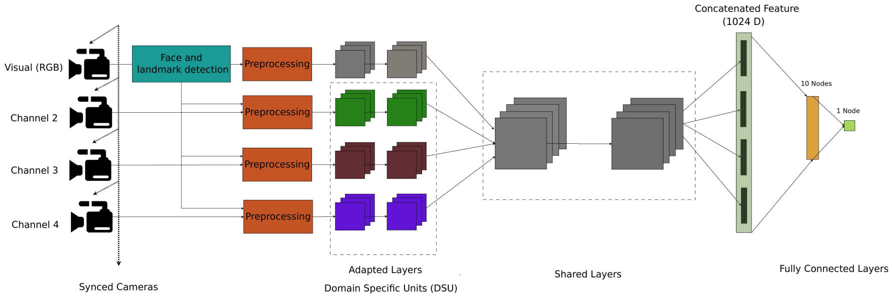


Fig. 1 Block diagram of the MC-CNN network. The gray color blocks in the CNN part represent layers which are not retrained, and other colored blocks represent re-trained/adapted layers. Note that the original approach from [29] is depicted here: it takes grayscale, infrared, depth and thermal data as input. The channels used can be changed depending of the available channels.

This architecture uses a late fusion strategy for combining multiple channels for the face PAD problem. The main idea in the Multi-Channel CNN (MC-CNN) is to use the joint representation from multiple modalities for PAD, using transfer learning from a pre-trained face recognition network [29]. The underlying hypothesis is that the joint representation in the face space could contain discriminative information for PAD. This network consists of three parts: low and high level convolutional/pooling layers, and fully connected layers, as shown in Figure 1. As noted in [53], high-level features in deep convolutional neural networks trained in the visual spectrum are domain-independent i.e. they do not depend on a specific modality. Consequently, they can be used to encode face images collected from different image sensing domains. The parameters of this CNN can then be split into higher-level layers (shared among the different channels), and lower-level layers (known as Domain Specific Units). 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. During training, low-level layers are adapted separately for different modalities, while shared higher-level layers remain unaltered. In the last part of the network, embeddings extracted from all modalities are concatenated, and two fully connected layers are added. The first fully connected layer has ten nodes, and the second one has one node. Sigmoidal activation functions are used in each fully connected layer, as in the original implementation [29]. These layers, added on top of the concatenated representations, are tuned exclusively for the PAD task using the Binary Cross Entropy as the loss function.

The MC-CNN approach hence introduces a novel solution for multimodal PAD problems, leveraging a pre-trained network for face recognition when a limited amount of data is available for training PAD systems. Note that this architecture can be easily extended for an arbitrary number of input channels.

Later, in [25] this work was extended to a one-class implementation utilizing a newly proposed one class contrastive loss (OCCL) and Gaussian mixture model. Essentially, the new loss function forces the network to learn a compact embedding for the bonafide channel, making sure that attacks are far from the bonafide attacks. This network learned is used as a fixed feature extractor and used together with a one-class Gaussian mixture model to perform the final classification. This approach yielded better results in unseen attacks.

3.2.2 Early Fusion: Multi-Channel Pixel-wise Binary Supervision(MC-PixBiS)

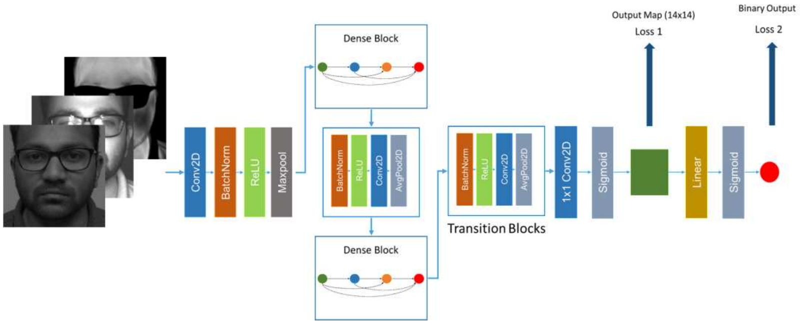


Fig. 2 MC-PixBiS architecture with pixel-wise supervision. Input channels are stacked before being passed to a series of dense blocks.

This architecture showcases the use of early fusion for a multi-channel PAD system. The Multi-Channel Deep Pixel-wise Binary Supervision network (MC-PixBiS) is a multi-channel extension of a recently published work on face PAD using legacy RGB sensors [23]. The main idea in [23] is to use pixel-wise supervision as an auxiliary supervision. The pixel-wise supervision forces the network to learn shared representations, and it acts as a patch-wise method (see Figure 2). To extend this network for a multimodal scenario, the method proposed in [66] was used, i.e., averaging the filters in the first layer and replicating the weights for different modalities.

The general block diagram of the framework is shown in Figure 2 and is based on DenseNet [33]. The first part of the network contains eight layers, and each layer consists of two dense blocks and two transition blocks. The dense blocks consist of dense connections between every layer with the same feature map size, and the transition blocks normalize and downsample the feature maps. The output from the eighth layer is a map of size 14 × 14 with 384 features. A 1 × 1 convolution layer is added along with sigmoid activation to produce the binary feature map. Further, a fully connected layer with sigmoid activation is added to produce the binary output. A combination of losses is used as the objective function to minimize:

$$\mathscr{L} = \lambda \mathscr{L}_{pix} + (1-\lambda)\mathscr{L}_{bin} \tag{1}$$

where $\mathscr{L}_{pix}$ is the binary cross-entropy loss applied to each element of the 14×14 binary output map and $\mathscr{L}_{bin}$ is the binary cross-entropy loss on the network's binary output. A λ value of 0.5 was used in our implementation. Even though both losses are used in training, in the evaluation phase, only the pixel-wise map is used: the mean value of the generated map is used as a score reflecting the probability of bonafide presentation.

3.2.3 Hybrid (Multi-head): Cross Modal Focal Loss (RGBD-MH(CMFL))

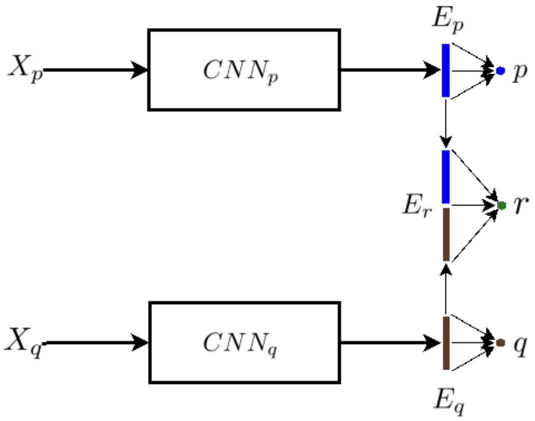


Fig. 3 Diagram of the two-stream multi-head model, showing the embeddings and probabilities from individual and joint branches. This can be extended to multiple heads as well.

The architecture presented here shows a hybrid approach to presentation attack detection [26]. A multi-head architecture that follows a hybrid fusion strategy is detailed here. The architecture of the network is shown in Fig. 4. Essentially, the architecture consists of a two-stream network with separate branches for the component channels. The embeddings from the two channels are combined to form the

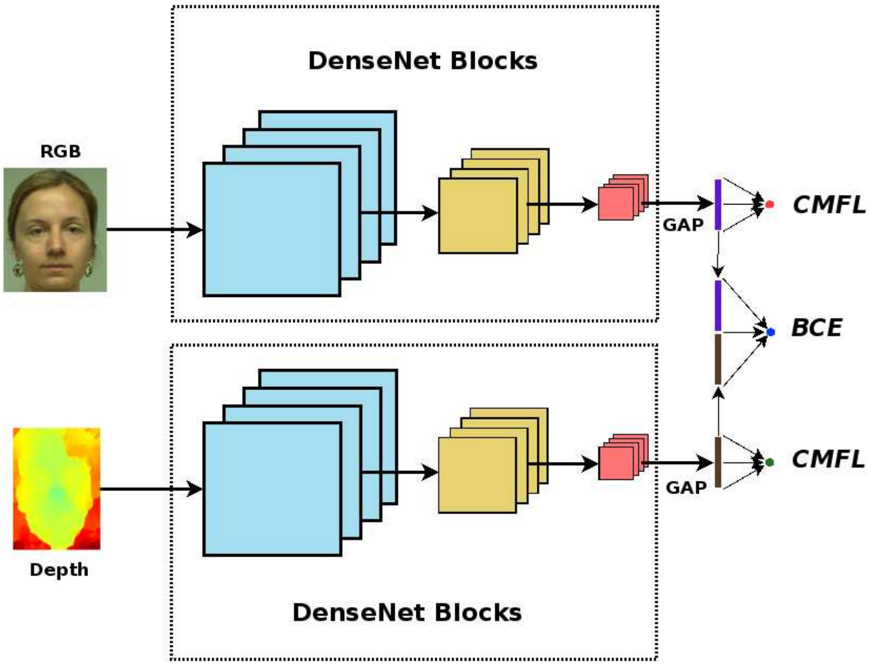


Fig. 4 The proposed framework for PAD. A two stream- multi-head architecture is used following a late fusion strategy. Heads corresponding to individual channels are supervised by the proposed cross-modal focal loss (CMFL), while the joint model is supervised by binary cross entropy (BCE).

third branch. Fully connected layers are added to each of these branches to form the final classification head. These three heads are jointly supervised by a loss function which forces the network to learn discriminative information from individual channels as well as the joint representation, reducing the possibility of overfitting. The multi-head structure also makes it possible to perform scoring even when a channel is missing at test time, meaning that we can do scoring with RGB branch alone (just using the score from the RGB head) even if the network was trained on a combination of two channels.

The individual branches are comprised of the first eight blocks (following the DeepPixBiS architecture [23]) from DenseNet architecture (densenet161) proposed by Huang *et al*. [34]. In the DenseNet architecture, each layer is connected to every other layer, reducing the vanishing gradient problem while reducing the number of parameters. We used pre-trained weights from the Image Net dataset to initialize the individual branches. The number of input channels for the RGB and depth channels has been modified to 3 and 1 for the RGB and depth channels, respectively. For the depth branch, the mean values of three-channel weights are used to initialize the weights of the modified convolutional kernels in the first layer. In each branch, a global average pooling (GAP) layer is added after the dense layers to obtain a 384-dimensional embedding. The RGB and depth embeddings are concatenated to form the joint embedding layer. A fully connected layer, followed by a sigmoid activation is added on top of each of these embeddings to form the different heads in the framework. At training time, each of these heads is supervised by a separate loss function. At test time, the score from the RGB-D branch is used as the PAD score.

A cross-modal focal loss (CMFL) to supervise the individual channels is also proposed in this work [26]. The core idea is that, when one of the channels can correctly classify a sample with high confidence, then the loss contribution of the sample in the other branch can be reduced. If a channel can correctly classify a sample confidently, then we don't want the other branch to penalize the model more. CMFL forces each branch to learn robust representations for individual channels, which can then be utilized with the joint branch, effectively acting as an auxiliary loss function.

The idea of relaxing the loss contribution of samples correctly classified is similar to the Focal Loss [41] used in object detection problems. In Focal Loss, a modulating factor is used to reduce the loss contributed by samples that are correctly classified with high confidence. We use this idea by modulating the loss factoring in the confidence of the sample in the current and the alternate branch.

Consider the two-stream multi-branch multi-head model in Fig. 3. X_p and X_q denotes the image inputs from different modalities, and E_p, E_q, and E_r denotes the corresponding embeddings for the individual and joint representations. In each branch, after the embedding layer, a fully connected layer (followed by a sigmoid layer) is present which provides classification probability. The variables p, q and r denote these probabilities.

The naive way to train a model is to use *BCE* loss on all three branches as:

$$\mathscr{L} = \mathscr{L}_p + \mathscr{L}_q + \mathscr{L}_r \tag{2}$$

Where each loss function is BCE. However, this approach penalizes all miss-classifications equally from both branches.

The Cross Modal Loss Function (CMFL) is given as follows:

$$CMFL p_{\mathrm{t}}, q_{\mathrm{t}} = -\alpha_{\mathrm{t}}(1 - w(p_t, q_t))^{\gamma} \log(p_{\mathrm{t}}) \tag{3}$$

The function $w(p_t, q_t)$, depends on the probabilities given by the channels from two individual branches. This modulating factor should increase as the probability of the other branch increases, and at the same time should be able to prevent very confident mistakes. The harmonic mean of both the branches weighted by the probability of the other branch is used as the modulating factor. This reduces the loss contribution when the other branch is giving confident predictions. And the expression for this function is given as:

$$w(p_t, q_t) = q_t \frac{2 p_t q_t}{p_t + q_t} \tag{4}$$

Note that the function w is asymetric, i.e., the expression for $w(q_t, p_t)$ is:

$$w(q_t, p_t) = p_t \frac{2p_t q_t}{p_t + q_t} \tag{5}$$

meaning the weight function depends on the probability of the other branch. Now we use the proposed loss function as auxiliary supervision,and the overall loss function to minimize is given as:

$$\mathscr{L} = (1-\lambda)\mathscr{L}_{\mathrm{CE}(r_{\mathrm{t}})} + \lambda(\mathscr{L}_{CMFL_{p_{\mathrm{t}},q_{\mathrm{t}}}} + \mathscr{L}_{CMFL_{q_{\mathrm{t}},p_{\mathrm{t}}}}) \tag{6}$$

The value of λ was non-optimally as 0.5 for the study. When the probability of the other branch is zero, then the loss is equivalent to standard cross-entropy. The loss contribution is reduced when the other branch can correctly classify the sample. i.e., when an attack example is misclassified by network CNN_p, the network CNN_p is penalized unless model CNN_q can classify the attack sample with high confidence. As the $w(p,q) \rightarrow 1$ the modulating factor goes to zero, meaning if one channel can classify it perfectly, then the other branch is less penalized. Also, the focussing parameter γ can be adapted to change the behavior of the loss curve. We used an empirically obtained value of $\gamma = 3$ in all our experiments.

4 Experiments

We have used the *HQ-WMCA* dataset for the experiments, which contains a wide variety of 2D, 3D, and partial attacks, collected from different channels such as color, thermal, infrared, depth, and short-wave infrared.

4.1 Dataset: HQ-WMCA

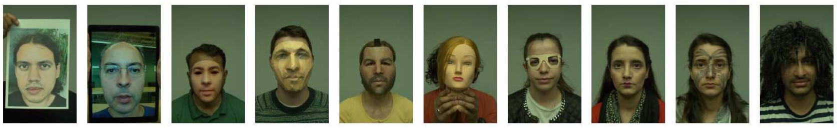


Fig. 5 Attacks present in *HQ-WMCA* dataset: (a) Print, (b) Replay, (c) Rigid mask, (d) Paper mask, (e) Flexible mask, (f) Mannequin, (g) Glasses, (h) Makeup, (i) Tattoo and (j) Wig. Image taken from [30].

The High-Quality Wide Multi-Channel Attack (*HQ-WMCA*)[30, 47] dataset consists of 2904 short multi-channel video recordings of both bonafide and presentation attacks. This database again consists of a wide variety of attacks including both obfuscation and impersonation attacks. Specifically, the attacks considered are print, replay, rigid mask, paper mask, flexible mask, mannequin, glasses, makeup, tattoo, and wig (Fig. 5). The database consists of recordings from 51 different subjects, with several channels including color, depth, thermal, infrared (spectra), and short-wave infrared (spectra). In this work, we consider the RGB channel captured with Basler acA1921-150uc camera and depth image captured with Intel RealSense D415.

4.2 Protocols

We use the grand_test as well as the leave-one-out (LOO) attack protocols distributed with the *HQ-WMCA* dataset. Specifically, in the LOO protocols, one attack is left out in the train and development set and the evaluation set consists of bonafide and the attack which was left out in the train and development set. This constitutes the unseen attack protocols or zero-shot attack protocols. The performance of the PAD methods in these protocols gives a more realistic estimate of their robustness against unseen attacks in real-world scenarios. In addition, we performed experiments with known attack protocols to evaluate the performance in a known attack scenario.

4.3 Metrics

For the evaluation of the algorithms, we have used the ISO/IEC 30107-3 metrics [35], Attack Presentation Classification Error Rate (APCER), and Bonafide Presentation Classification Error Rate (BPCER) along with the Average Classification Error Rate (ACER) in the *eval* set. We compute the threshold in the *dev* set for a BPCER value of 1%, and this threshold is applied in the *eval* set to compute the reported metrics.

$$ACER = \frac{APCER + BPCER}{2}. \tag{7}$$

4.4 Implementation details

We performed data augmentation during the training phase with random horizontal flips with a probability of 0.5. The combined loss function is minimized with Adam Optimizer [38]. A learning rate of 1×10^{-4} was used with a weight decay parameter of 1×10^{-5}. We used a mini-batch size of 64, and the network was trained for 25 epochs on a GPU grid. The architecture and the training framework were implemented using the PyTorch [51] library.

4.5 Baselines

For a fair comparison with state-of-the-art, we have implemented 3 different multi-channel PAD approaches from literature as described in section 3.2 for the RGB-D channels. Besides, we also introduce a multi-head architecture supervised with BCE alone, as another baseline for comparison. The baselines implemented are listed below.

RGB-DeepPixBiS: This is an RGB only CNN based system [23], trained using both binary and pixel-wise binary loss function. This model is used as a baseline for comparing with multi-channel models.

MC-PixBiS: This is a CNN based system [23], extended to multi-channel scenario as described in [30] trained using both binary and pixel-wise binary loss function. This model uses RGB and depth channels stacked together at the input level.

MCCNN-OCCL-GMM: This model is the multi-channel CNN system proposed to learn one class model using the one class contrastive loss (OCCL) and Gaussian mixture model as reported in [25]. The model was adapted to accept RGB-D channels as the input.

MC-ResNetDLAS: This is the reimplementation of the architecture from [50], which won the first prize in the ‘CASIA-SURF’ challenge, extending it to RGB-D channels, based on the open-source implementation [2]. We used the initialization from the best-pretrained model as suggested in [50] followed by retraining in the current protocols using RGB-D channels.

RGBD-MH-BCE: This uses the multi-head architecture shown in Fig.4, where all the branches are supervised by binary cross-entropy (BCE). In essence, this is equivalent to setting the value of $\gamma = 0$, in the expression for the cross-modal loss function. This is shown as a baseline to showcase the improvement by the new multi-head architecture alone and to contrast with the performance change with the new loss function.

RGBD-MH-CMFL: This uses the multi-head architecture shown in Fig.4, where individual branches are supervised by CMFL loss and joint branch is supervised by BCE loss.

4.6 Experiments and Results

Results in HQ-WMCA dataset: Table 1 shows the performance of different methods in the grandtest protocol, which evaluates the performance of the methods in a known attack scenario, meaning all attacks are distributed equally across train, development, and test set. From the ACER values, it can be seen that the multi-head architecture performs the best followed by MC-PixBiS architecture. The RGB alone model (RGB-DeepPixBiS) also performs reasonably well in this protocol. The corresponding ROC plots for the evaluation set are shown in Fig. 6. It can be seen from the ROC that the RGB-only model outperforms the multi-head model in the evaluation set opposite to the results in Table 1. In ACER evaluations, the threshold used

Table 1 Performance of the multi-channel systems in the grandtest protocol of HQ-WMCA dataset. The values reported are obtained with a threshold computed for BPCER 1% in *dev* set.

	APCER	BPCER	ACER
RGB-DeepPixBiS	9.2	0.0	4.6
MC-PixBiS	9.7	0.0	4.8
MCCNN-OCCL-GMM	7.9	11.4	9.7
MC-ResNetDLAS	8.0	6.4	7.2
RGBD-MH-BCE	4.0	2.0	3.0
RGBD-MH-CMFL	6.6	0.1	3.3

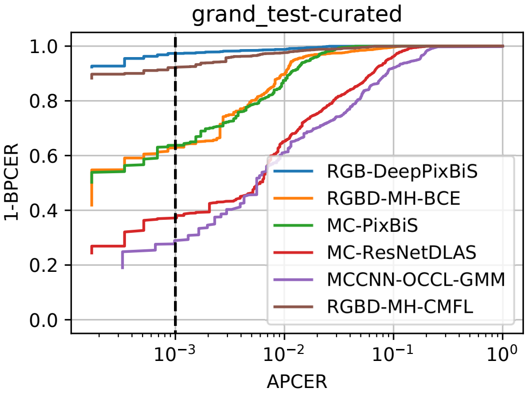


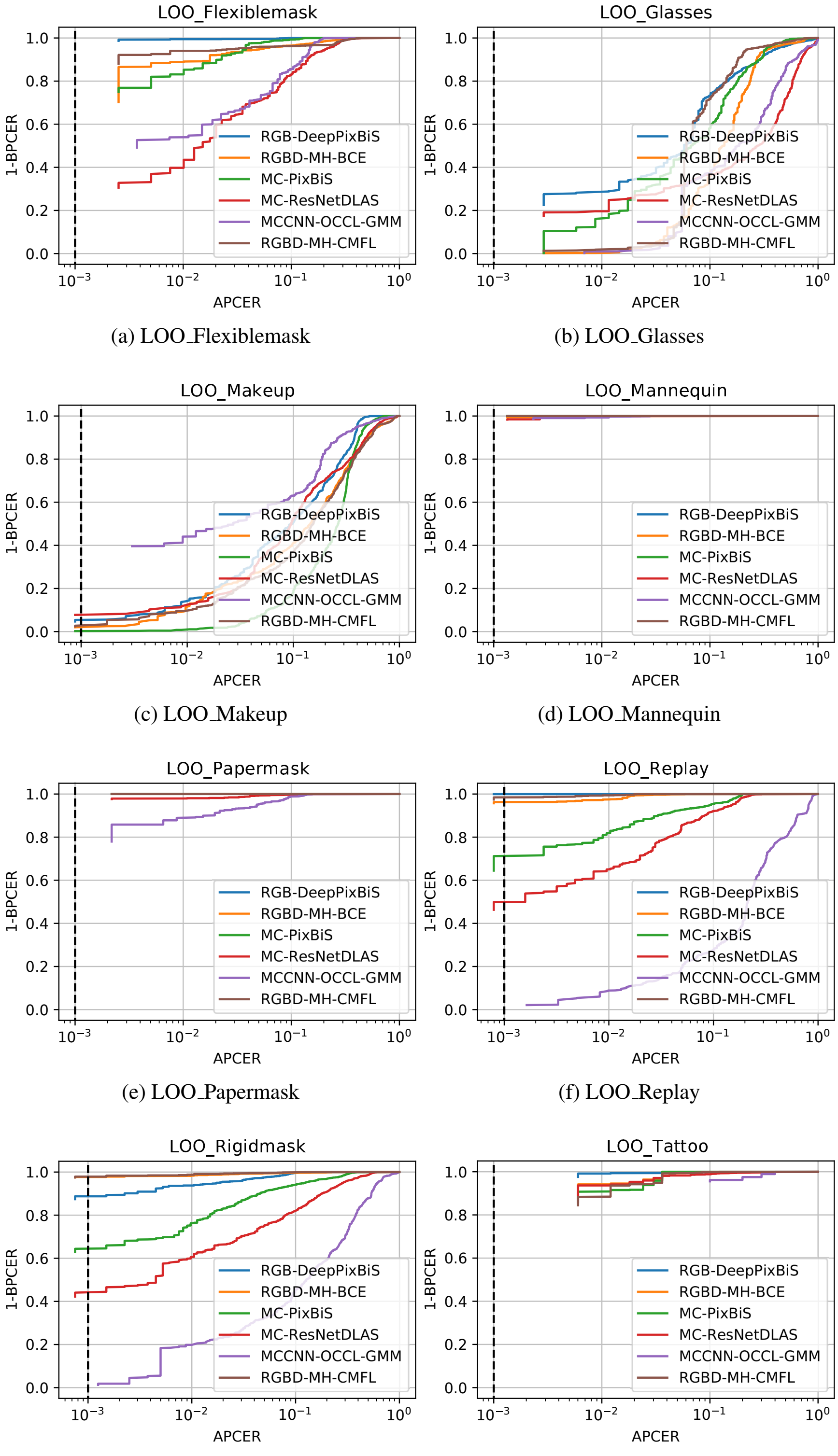
Fig. 6 ROC plot for evaluation set in the grandtest protocol in HQ-WMCA dataset.

Table 2 Performance of the multi-channel systems method in **unseen** protocols of HQ-WMCA dataset. The values reported are obtained with a threshold computed for BPCER 1% in *dev* set.

	Flexiblemask	Glasses	Makeup	Mannequin	Papermask	Rigidmask	Tattoo	Replay	Mean±Std
RGB-DeepPixBiS [23]	5.8	49.3	23.8	0.0	0.0	25.9	13.6	6.0	15.5± 15.8
MC-PixBiS [23]	29.9	49.9	29.4	0.1	0.0	32.5	5.7	9.6	19.6±17.1
MCCNN-OCCL-GMM [25]	14.2	32.7	22.0	1.5	7.1	33.7	4.2	36.6	19.0±13.2
MC-ResNetDLAS [50]	23.5	50.0	33.8	1.0	2.6	31.0	5.7	15.5	20.3±16.2
RGBD-MH-BCE	16.7	38.1	43.3	0.4	1.3	3.0	2.0	2.3	13.3±16.5
RGBD-MH-CMFL [26]	14.8	37.4	34.9	0.0	0.4	2.4	2.4	1.0	**11.6±14.8**

is selected from the development set. The ROC plots only depict the performance in the evaluation set without considering the threshold selected from the development set causing the discrepancy. ACER reported shows more realistic performance estimates in real-world scenarios as the thresholds are fixed in advance according to specific performance criteria.

The *HQ-WMCA* dataset consists of challenging attacks, specifically, there are different types of partial attacks such as *Glasses* which occupy only a part of the face. These attacks are much harder to detect when they are not seen in the training set, as they appear very similar to bonafide samples. The analysis we performed is similar to a worst-case analysis since it specifically focuses on the unseen attack robustness. The experimental results in the LOO protocols of *HQ-WMCA* are tabulated in Table 2. Overall, *MCCNN-OCCL-GMM* and *MC-ResNetDLAS* do not perform well



(a) LOO_Flexiblemask

(b) LOO_Glasses

(c) LOO_Makeup

(d) LOO_Mannequin

(e) LOO_Papermask

(f) LOO_Replay

(g) LOO_Rigidmask

(h) LOO_Tattoo

Fig. 7 ROC plots for the evaluation set in **unseen attack** protocols of HQ-WMCA dataset.

in the LOO protocols of *HQ-WMCA* database. In addition, the *MC-PixBiS* method also performs poorly in the unseen attack protocols of the *HQ-WMCA* dataset. This could be due to the challenging nature of the attacks in the database. The RGB-only method *RGB-DeepPixBiS* performs reasonably well overall too. It can be seen that the multi-head architecture, *RGBD-MH-BCE*, already improves the results as compared to all the baselines with an average ACER of 13.3 ± 16.5. With the addition of the *CMFL* loss, the ACER further improves to $11.6 \pm 14.8\%$. The results indicate that the proposed architecture already improves the performance in challenging attacks, and the proposed loss further improves the results achieving state-of-the-art results in the *HQ-WMCA* dataset.

From Table 2, it can be seen that the effectiveness of fusion strategies is different for unseen PA scenarios. While most of the multi-channel methods struggle to achieve good performance in detecting unseen *Flexiblemasks*, the *RGB-DeepPixBiS* achieves much better performance in this case. A similar trend can be seen in the case of *Makeup* attack as well. This could be due to the lack of additional information provided by the depth channel in these cases. The depth information in the case of these attacks is very similar to that of bonafide samples. However, multi-channel method provides a significant boost in performance in detecting *Rigidmask*, *Tattoo* and *Replay* attacks. Attacks like *Papermask* and *Mannequins* are easier to detect in most of the cases due to the distinct appearance compared to bonafide samples. The multi-head architecture improves the performance compared to other baselines in most of the sub-protocols. The ROC plots for the eval set for the corresponding protocols are shown in Fig. 7.

Performance with missing channels: We evaluate the performance of the multi-head models when evaluated with only a single channel at test time. Consider a scenario where the model was trained with RGB and depth, and at the test time, only one of the channels is available. We compare with the mean performance in the *HQ-WMCA* dataset, with RGB and depth alone at test time. The results are shown in Table 3. For the baseline *RGBD-MH-BCE*, using RGB alone at test time the error rate is 15.4 ± 16.1, whereas for the proposed approach it improves to 12.0 ± 13.9. The performance improves for the depth channel as well.

From Table 3, it can be seen that the performance improves, as compared to using BCE even when using a single channel at the time of deployment. This shows that the performance of the system improves when the loss contributions of samples that are not possible to classify by that modality are reduced. Forcing the individual networks to learn a decision boundary leads to overfitting resulting in poor generalization.

Table 3 Ablation study using only one channel at deployment time.

	RGB	Depth
RGBD-MH-BCE	15.4±16.1	34.2±11.6
RGBD-MH-CMFL	**12.0±13.9**	**30.6±17.5**

4.7 Computational Complexity

Table 4 Computational and parameter complexity comparison

Model	Compute	Parameters
MCCNN-OCCL-GMM [25]	14.51 GMac	50.3 M
MC-PixBiS [23]	4.7 GMac	3.2M
MC-ResNetDLAS [50]	15.34 GMac	69.29 M
RGBD-MH(CMFL) [26]	9.16 GMac	6.39 M

Here we compare the complexity of the models in terms of parameters and compute required (for RGB and Depth channels). The comparison is shown in Table 4. It can be seen that the parameter and compute for late fusion (MCCNN-OCCL-GMM) is quite high. A lot of additional parameters are added for each channel before fusion which increases the total number of parameters. The MC-ResNetDLAS also suffers from a high number of parameters and compute. The Early fusion method, MC-PixBiS, with the truncated DenseNet architecture saves compute and parameters a lot compared to others. Thanks to fusing the channels at the input level, the parameter increase is just for the first convolutional filter keeping rest of the operations the same. This makes it easy to add more channels as the rest of the network remains the same except for the first convolutional filter. Lastly, the RGBD-MH(CMFL) is composed of the PixBiS model for each of the channels, and hence roughly double the number of parameters and compute compared to the PixBiS model.

4.8 Discussions

From the results, it was observed that the late fusion method MCCNN-OCCL-GMM performed poorly compared to other methods. Also, this strategy increases the number of parameters with the increase in the number of channels. The MC-PixBiS model, on the other hand, does not increase the number of parameters with the increase in the number of channels. Each additional channel only changes the parameters in the first convolutional filter, which is negligible compared to the total number of parameters. In short, the early fusion method is more scalable to an increasing number of channels as the computational complexity is very less. However, this method cannot handle a missing channel scenario in a real-world application.

From the performance evaluations, it was seen that the RGBD-MH(CMFL) architecture achieves the best performance. Usage of multiple heads forces the network to learn multiple redundant features from individual channels as well as from the joint representation. This effectively acts as a regularization mechanism for the network, preventing overfitting to seen attacks. Further, one limitation of the multi-head representation is that the network is forced to lea discriminative features from all the channels, which may not be trivial. The CMFL loss proposed effectively addresses this issue by dynamically modulating the loss contribution from individual channels. Comparing the computational complexity, this model is relatively simpler compared to the late fusion model. Nevertheless, this model is more complex compared to the early fusion approach with nearly double the number of parameters, however, with the addition of parameters and the formulation it can be seen that the architecture itself improves the robustness, and again with the use of CMFL loss, the performance further improves indicating a good performance complexity tradeoff. The cross-modal focal loss function modulates the loss contribution of samples based on the confidence of individual channels. The framework can be trivially extended to multiple channels and different classification problems where information from one channel alone is inadequate for classification. This loss forces the network to learn complementary, discriminative, and robust representations for the component channels. The structure of the framework makes it possible to train models using all the available channels and to deploy with a subset of channels. One limitation of the framework is that the addition of more channels requires more branches which increases the parameters linearly with the number of channels. While we have selected RGB and depth channels for this study, mainly due to the availability of off-the-shelf devices consisting of these channels, it is trivial to extend this study to other combinations of channels as well, for instance, RGB-Infrared, and RGB-Thermal.

5 Conclusions

In this chapter, we have presented different approaches using multi-channel information for presentation attack detection. All the approaches have their merits and limitations, however, we have conducted an extensive analysis of the unseen attack robustness as a worst-case performance evaluation. As multi-channel methods are required for safety-critical applications, robustness against unseen attacks is an essential requirement. In the evaluations, we have noted that the performance is much better for the multi-head architecture, thanks to the CMFL loss. The CMFL loss forces the network to learn complementary, discriminative, and robust representations for the component channels. This work can be straightforwardly extended to other combinations of channels and architectures as well. The *HQ-WMCA* database and the source code and protocols will be made available to download for research purposes. This will certainly foster further research in multi-channel face presentation attack detection in the future.

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.

Index

Glossary

ACER: Average Classification Error Rate , **16**
APCER: Attack Presentation Classification Error Rate , **16**

BCE: Binary Cross Entropy , **15**
BPCER: Bonafide Presentation Classification Error Rate , **16**
BSIF: Binarized statistical image features, **3**

CMFL: Cross Modal Focal Loss , **15**
CNN: Convolutional Neural Network, **1**, **3**

HQ-WMCA: High-Quality Wide Multi-Channel Attack , **16**
HTER: Half Total Error Rate, **3**

IQM: Image Quality Measures, **3**

LOO: Leave-one-out , **16**

NIR: Near-Infrared, **1**

PA: Presentation Attack, **1**
PAD: Presentation Attack Detection, **1**
PAD: Presentation Attack Detection , **4**
PixBiS: Pixel-wise Binary Supervision , **21**

RBF: Radial Basis Function, **3**
RNN: Recurrent Neural Network, **3**

SVM: Support Vector Machines, **3**
SWIR: Short-wave Infrared, **1**

References

1. About face id advanced technology (2021). URL https://support.apple.com/en-gb/HT208108
2. ResNetDLAS. https://github.com/AlexanderParkin/ChaLearn_liveness_challenge/ (2021). [Online; accessed 1-Feb-2021]
3. Agarwal, A., Yadav, D., Kohli, N., Singh, R., Vatsa, M., Noore, A.: Face presentation attack with latex masks in multispectral videos. SMAD **13**, 130 (2017)
4. 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)
5. Arashloo, S.R., Kittler, J., Christmas, W.: An anomaly detection approach to face spoofing detection: A new formulation and evaluation protocol. IEEE Access **5**, 13,868–13,882 (2017)

6. 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)
7. 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)
8. Bhattacharjee, S., Mohammadi, A., Marcel, S.: Spoofing deep face recognition with custom silicone masks. In: 2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS), pp. 1–7. IEEE (2018)
9. Bhattacharjee, S., Mohammadi, A., Marcel, S.: Spoofing deep face recognition with custom silicone masks. In: 2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS), pp. 1–7. IEEE (2018)
10. 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. In: 2017 IEEE International Joint Conference on Biometrics (IJCB), pp. 688–696. IEEE (2017)
11. 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)
12. 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)
13. 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)
14. 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)
15. 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)
16. Engelsma, J.J., Jain, A.K.: Generalizing fingerprint spoof detector: Learning a one-class classifier. arXiv preprint arXiv:1901.03918 (2019)
17. Erdogmus, N., Marcel, S.: Spoofing face recognition with 3d masks. IEEE transactions on information forensics and security **9**(7), 1084–1097 (2014)
18. 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)
19. de Freitas Pereira, T., Anjos, A., De Martino, J.M., Marcel, S.: Can face anti-spoofing countermeasures work in a real world scenario? In: 2013 international conference on biometrics (ICB), pp. 1–8. IEEE (2013)
20. 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)
21. 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)
22. George, A., Geissbuhler, D., Marcel, S.: A comprehensive evaluation on multi-channel biometric face presentation attack detection. arXiv preprint arXiv:2202.10286 (2022)
23. George, A., Marcel, S.: Deep pixel-wise binary supervision for face presentation attack detection. International Conference on Biometrics (2019)
24. 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)

25. 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)
26. George, A., Marcel, S.: Cross modal focal loss for RGBD face anti-spoofing. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
27. George, A., Marcel, S.: Multi-channel Face Presentation Attack Detection Using Deep Learning, pp. 269–304. Springer International Publishing, Cham (2021). DOI 10.1007/978-3-030-74697-1_13
28. George, A., Marcel, S.: On the effectiveness of vision transformers for zero-shot face anti-spoofing. In: 2021 IEEE International Joint Conference on Biometrics (IJCB), pp. 1–8. IEEE (2021)
29. 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
30. 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)
31. Heusch, G., Marcel, S.: Pulse-based features for face presentation attack detection (2018)
32. Heusch, G., Marcel, S.: Remote blood pulse analysis for face presentation attack detection. In: Handbook of Biometric Anti-Spoofing, pp. 267–289. Springer (2019)
33. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely Connected Convolutional Networks. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), vol. 1, p. 3 (2017)
34. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: CVPR, vol. 1, p. 3 (2017)
35. Information technology –International Organization for Standardization. Standard, International Organization for Standardization (2016)
36. Jaiswal, A., Xia, S., Masi, I., AbdAlmageed, W.: Ropad: Robust presentation attack detection through unsupervised adversarial invariance. arXiv preprint arXiv:1903.03691 (2019)
37. Kaichi, T., Ozasa, Y.: A hyperspectral approach for unsupervised spoof detection with intra-sample distribution. In: 2021 IEEE International Conference on Image Processing (ICIP), pp. 839–843. IEEE (2021)
38. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (Poster) (2015)
39. 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)
40. Li, L., Xia, Z., Li, L., Jiang, X., Feng, X., Roli, F.: Face anti-spoofing via hybrid convolutional neural network. In: the Frontiers and Advances in Data Science (FADS), 2017 International Conference on, pp. 120–124. IEEE (2017)
41. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection (2018)
42. 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)
43. Liu, Y., Stehouwer, J., Jourabloo, A., Liu, X.: Deep tree learning for zero-shot face anti-spoofing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4680–4689 (2019)
44. 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)
45. 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)

46. Mehta, S., Uberoi, A., Agarwal, A., Vatsa, M., Singh, R.: Crafting a panoptic face presentation attack detector
47. Mostaani, Z., , A., Heusch, G., Geissenbuhler, D., Marcel, S.: The high-quality wide multi-channel attack (hq-wmca) database (2020)
48. 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)
49. 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)
50. 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)
51. 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)
52. 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)
53. Pereira, T.d.F., Anjos, A., Marcel, S.: Heterogeneous Face Recognition Using Domain Specific Units. IEEE Trans. on Information Forensics and Security (TIFS) **14**(7), 1803–1816 (2019)
54. Perera, P., Patel, V.M.: Learning deep features for one-class classification. IEEE Transactions on Image Processing **28**(11), 5450–5463 (2019)
55. 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)
56. Purnapatra, S., Smalt, N., Bahmani, K., Das, P., Yambay, D., Mohammadi, A., George, A., Bourlai, T., Marcel, S., Schuckers, S., et al.: Face liveness detection competition (livdet-face)-2021. In: 2021 IEEE International Joint Conference on Biometrics (IJCB), pp. 1–10. IEEE (2021)
57. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 (2015)
58. Raghavendra, R., Raja, K.B., Busch, C.: Presentation attack detection for face recognition using light field camera. IEEE Transactions on Image Processing **24**(3), 1060–1075 (2015)
59. 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)
60. Ramachandra, R., Busch, C.: Presentation attack detection methods for face recognition systems: a comprehensive survey. ACM Computing Surveys (CSUR) **50**(1), 8 (2017)
61. Sepas-Moghaddam, A., Pereira, F., Correia, P.L.: Light field-based face presentation attack detection: Reviewing, benchmarking and one step further. IEEE Transactions on Information Forensics and Security **13**(7), 1696–1709 (2018)
62. 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)
63. Siegmund, D., Kerckhoff, F., Magdaleno, J.Y., Jansen, N., Kirchbuchner, F., Kuijper, A.: Face presentation attack detection in ultraviolet spectrum via local and global features. In: 2020 International Conference of the Biometrics Special Interest Group (BIOSIG), pp. 1–5. IEEE (2020)
64. 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)
65. 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)

66. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. In: European Conf. on Computer Vision (ECCV), pp. 20–36. Springer (2016)
67. 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)
68. 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)
69. Yang, J., Lei, Z., Li, S.Z.: Learn convolutional neural network for face anti-spoofing. arXiv preprint arXiv:1408.5601 (2014)
70. 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)
71. 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)