Autoencoders for Open-Set Presentation Attack Detection

Igor Bastos, *Member, IEEE,* Anjith George, *Member, IEEE,*
Sebastien Marcel, *Senior Member, IEEE,* Anderson Rocha, *Fellow, IEEE,*

***Abstract*—Face anti-spoofing (FAS) has gained significant attention due to its crucial role in ensuring the security of facial recognition systems. Despite numerous proposed methods for addressing various types of presentation attacks, there remains a gap in the existing literature, as most methods only tackle seen attack types and lack generalizability in real-world scenarios. This paper presents a novel method for face antispoofing, specifically focused on open-set scenarios. The proposed approach consists of an autoencoder exclusively trained to reconstruct bona fide data. The key idea is that features extracted using this model would exhibit poor quality for attack data, resulting in significantly higher reconstruction errors. Consequently, these low-quality features would differ from the well-reconstructed features of bona fide samples. Our method's key novelty lies in the rich feature learning procedure that focuses on the properties of bona fide samples through an autoencoder formulation. The extracted features are used to train one-class classifiers using only bona fide samples. Experimental results on standard benchmarks such as RECOD-MPAD, WMCA, HQ-WMCA, Replay-Attack, OULU-NPU and MSU-MFSD datasets demonstrate the excellent performance of our proposed approach, placing it among state-of-the-art methods for metrics like Bona Fide Presentation Classification Error Rate (BPCER), Attack Presentation Classification Error Rate (APCER) and Average Classification Error Rate (ACER).**

***Index Terms*—Open-set classification, Autoencoders, Face Anti-Spoofing.**

I. Introduction

THE advancements in camera and sensor technology have led to the development of highly sophisticated face recognition (FR) systems, which have been successful in several applications, being the subject of extensive research in the past years [30], [9], [38], [4]. However, despite their good results, FR systems remain vulnerable to Presentation Attacks (PAs) [30], which attempt to deceive recognition mechanisms using fake or altered representations of genuine data. In face anti-spoofing, PAs involve photographs, masks, videos, or other methods to mimic real faces, compromising security and accuracy of face verification technologies [30], [9], [38], [4]. To counter these threats, various face anti-spoofing techniques have been developed, incorporating multimodal inputs and advanced deep neural networks to enhance resilience against various types of PA [50].

Regardless of significant advances in recent years and the proposal of increasingly complex methods, many face anti-spoofing (FAS) approaches struggle with unseen presentation attacks, leading to poor generalization [50]. This limitation weakens their robustness in real-world applications, where a malicious attacker can potentially deceive a face recognition (FR) system with a type of input not accounted for the set of trained PAs [30], [50]. In real-world scenarios, one can assume that all presentation attacks are unseen, as it is not possible to foretell the variations an FR system could encounter a priori. Consequently, an ideal approach would be to develop a FAS system that considers this issue, aligned with methods that employ the open-set problem formulation and only use the bona fide data during the training.

To address the aforementioned issue, this study presents AOPAD (Autoencoders for Open-Set Presentation Attack Detection), a method that builds upon a previously proposed gesture recognition method named MORA [5]. In contrast to traditional approaches, AOPAD utilizes a single autoencoder model that is exclusively trained using bona fide data. The underlying concept revolves around the notion that features obtained from this model would demonstrate lower reconstruction quality when applied to attack data. This can be observed by such samples' significantly higher reconstruction error. Hence, the low-quality features extracted from attack data would diverge from the well-reconstructed features of genuine samples.

To classify samples, features extracted from the trained autoencoder are used to train a combination of clustering algorithms and one-class classifiers to capture spread patterns of bona fide data and distinguish between bona fide samples and attacks. These features are derived from a task-specific reconstruction objective, where the model learns to predict multiple handcrafted descriptors, BSIF, SSR, LBP, and SWIR, from a single RGB or grayscale input. This unified strategy enriches the latent space with high-level cues related to texture, illumination, and material properties, enhancing the model's ability to detect spoofs. Thus, our method operates without relying on any information from presentation attacks, and even the hyperparameters are computed with no supervision, eliminating the need for validation data from attack classes. In order to evaluate the performance of AOPAD, extensive experiments were conducted on RECOD-MPAD [2], WMCA [16], HQ-WMCA [22], Replay-Attack [12], OULU-NPU [7] and MSU-MFSD [47], demonstrating that our method rivals with state-of-the-art approaches across these databases.

The key contributions of the present paper are listed below:

- A novel open-set framework for face anti-spoofing that operates on RGB or multimodal bona fide data, with no reliance on presentation attack samples, neither for training, validation, nor hyperparameter tuning. The framework combines a task-specific autoencoder with a cluster-aware one-class classification strategy, resulting in a

scalable and practical solution for real-world deployment.
- A unified feature reconstruction strategy where the autoencoder learns to generate high-level, handcrafted descriptors, such as BSIF, SSR, LBP, and SWIR, from a single input. This joint reconstruction enhances the latent space with complementary cues related to texture, reflectance, and material properties, offering a rich and discriminative representation for detecting spoofing attempts.
- A novel architectural design featuring an asymmetric encoder-decoder structure tailored for the anti-spoofing task. The encoder leverages a lightweight MobileVIT [35] backbone, combining the global modeling capacity of transformers with the efficiency of convolutional operations. The decoder consists of convolutional layers specialized in reconstructing spatially meaningful features. Departing from conventional autoencoders that aim to replicate the input, our model learns to reconstruct distinct anti-spoofing descriptors, resulting in intermediate activations that are rich in semantic cues and highly informative for spoof detection.
- A cluster-wise modeling approach that replaces the conventional single one-class classifier with multiple Isolation Forests trained on clusters of intermediate decoder activations. Clusters are discovered via PCA and K-means, with hyperparameters, such as the number of clusters and contamination rate, determined using silhouette scores, leading to better handling of intra-class variability without manual tuning.
- An extensive evaluation across diverse datasets and protocols, including RECOD-MPAD, WMCA, HQ-WMCA, Replay-Attack, OULU-NPU and MSU-MFSD, demonstrating that our method not only generalizes well to unseen attacks but also performs competitively with state-of-the-art supervised and open-set baselines under realistic conditions.

The source code to reproduce the experiments is available and accessible in $github.com/igorcrexito/spoofing\_reconstruction/$.
The remainder of this paper is organized as follows. Section 2 explores the literature, discussing concepts of related approaches that tackle the face anti-spoofing task, especially those that handle unseen attacks. Section 3 presents our approach, detailing all the steps of the framework. Section 4 presents the method evaluation, describing the datasets, results, and ablation studies for the experiments. In addition, this section also brings analysis of the obtained results. Finally, Section 5 concludes the work and outlines possible future research directions.

II. Literature Review

Majority of Facial Anti-Spoofing (FAS) techniques in the literature are focused on developing and utilizing deep neural models [50]. To achieve more accurate results and robustness to several types of presentation attacks, researchers propose increasingly complex architectures while incorporating innovative strategies like attention mechanisms [10], [23], [33], zero-sample learning [30], [43], and adversarial learning [31].

Although recent Facial Anti-Spoofing (FAS) methods show promising results, most treat it as a conventional classification problem [30], [50], training models on bona fide and known attack samples. This limits their practicality in real-world scenarios, reducing generalization to unseen attacks [30]. Recognizing this challenge, effective solutions must account for unseen PAs in FAS approaches. This review focuses on studies that highlight the importance of developing FAS methods resilient to unseen PAs.

The method introduced by Arashloo et al. [3] stands out as an early endeavor to tackle the issue of unseen presentation attacks. Their approach involves treating unseen PAs as anomalies and employing one-class Support Vector Machines (SVMs) to detect them. Another similar approach was proposed by Nikisins et al. [39], who utilized Gaussian Mixture Models (GMM) as an alternative anomaly detector instead of one-class SVMs. Similarly, Xiong and AbdAlmageed [48] utilized autoencoders in their experimentation, employing Local Binary Patterns (LBP) as input features and using both autoencoders and one-class SVMs as anomaly detectors. Although these methods are considered preliminary in addressing unseen attacks, they highlight the effectiveness of one-class classifiers in this scenario and their potential suitability in real-world situations. However, these approaches rely solely on the utilization of handcrafted texture features to extract information from inputs and do not take into account the distribution of bona fide data. This deficiency can lead to supplying very sparse data to the one-class classifiers, potentially compromising their ability to effectively generalize and detect anomalies.

Various methods aim to improve generalization to unseen attacks and distinguish genuine from attack samples. Baweja et al. [9] propose a pipeline that trains a one-class classifier while refining feature representations. Their approach generates pseudo-negative features from genuine data to simulate PAs, which are then used for training. Though it avoids real attack data, it relies on deformations applied to bona fide samples using a Gaussian distribution estimator. The method's success depends on accurately estimating this distribution. Consequently, if the distribution of synthetic attacked images substantially deviates from the assumed model, the generated pseudo-negative samples might inadequately capture the nuances of attacks, thus limiting the model's ability to generalize proficiently.

Some methodologies adopt zero or few-shot learning strategies to address the challenge of encountering unseen presentation attacks (PAs). For example, Liu et al. [30] introduced a zero-shot methodology harnessing a Deep Tree Network (DTN) to generate a semantic topology, without needing prior knowledge of unknown attacks. Similarly, Nguyen et al. [38] proposed an approach using an Adversarial Distribution Generative Network (ADGN) to craft a distinctive latent space, adeptly encompassing rare, unseen, and unfamiliar attacks [17], [28]. Despite yielding promising outcomes, zero or few-shot methodologies often struggle to generalize across diverse data types, a common challenge in handling unseen presentation attacks in real-world scenarios. Additionally, these methods typically rely on transferring knowledge

from seen to unseen classes. In face anti-spoofing, this transfer may not capture the intricacies of spoofing attacks, leading to poor generalization on unseen attack types. On the other hand, an alternative approach could exclusively focus on efficiently modeling of bona fide data and task-specific features, without depending on transfer learning or prior knowledge of unknown attacks. Consequently, the diversity and variability of unseen presentation attacks have minimal impact on the performance of this alternative method.

In turn, the method proposed by Arashloo [4] introduces a localized multiple kernel learning (MKL) technique tailored specifically for pure one-class classification of unknown or unseen face presentation attacks. This method successfully captures the intrinsic variability and local structure inherent in bona fide samples through a localized formulation, thus significantly enhancing detection performance. The methodology employs a strategy to organize data into clusters. Nonetheless, it does not offer a mechanism for determining the optimal number of clusters, an aspect notably absent from this approach. This limitation becomes conspicuous as the method exhibits a high sensitivity to parameter choices, including the number of clusters and the selection of kernel functions. In scenarios with certain bona fide data distributions, the absence of this selection mechanism could potentially deteriorate the performance of the method.

A recent method that also addresses unseen presentation attacks is Hyp-OC, proposed by Narayan and Patel [36]. It introduces a hyperbolic one-class classification framework that maps bona fide features into a Poincaré ball, improving compactness and separability, while generating pseudo-negatives through Gaussian sampling. Despite strong performance, Hyp-OC remains sensitive to hyperbolic optimization and may struggle when synthetic negatives fail to approximate real attacks. Complementary to this line of work, Liu et al. [33] proposed FM-VIT, a flexible multi-modal Vision Transformer that learns modality-agnostic liveness cues through cross-modal attention, enabling multi-modal training with single-modal deployment. Building on this direction, Liu et al. [34] introduced a unified FAS framework grounded in hierarchical prompt tuning, leveraging a Visual Prompt Tree to handle both physical presentation attacks and digital forgeries. Together, these approaches illustrate emerging trends that emphasize multi-modal feature integration and the use of semantic prompting to enhance generalization across diverse and previously unseen attack scenarios.

Recent work in face anti-spoofing has explored the use of vision-language models (VLMs) to enhance domain generalization. CFPL-FAS [32] introduces class-free prompt learning, leveraging content- and style-conditioned prompts derived through lightweight transformers to guide CLIP features toward more discriminative representations. Similarly, CCPE [21] proposes content-aware composite prompt engineering, combining instruction-based prompts from large language models with learnable visual prompts, further refined by a cross-modal guidance module that adaptively fuses semantic and visual information. These approaches demonstrate that integrating multimodal cues can improve robustness across diverse domains and attack scenarios. In contrast, AOPAD adopts an open-set paradigm, focusing exclusively on modeling bona fide data to capture the genuine feature distribution and treating attacks as out-of-distribution samples at inference time. By emphasizing simplicity and generality, AOPAD eliminates reliance on additional priors or attack examples, offering a practical solution for scenarios where unknown or scarce attack data is encountered.

Facial anti-spoofing (FAS) has made notable progress in mitigating presentation attacks (PAs), yet the majority of existing methods remain heavily dependent on PA data during training, limiting their capacity to generalize to unseen attack types. Typically framed as a binary classification problem, these approaches are trained on both bona fide and known attack samples, which hinders their effectiveness when confronted with novel or sophisticated spoofing techniques. Even methods that explicitly target the challenge of unseen attacks often resort to generating pseudo-attack samples or adopt complex architectures that introduce significant practical limitations. These solutions frequently rely on strong assumptions about the data distribution or demand extensive parameter tuning, reducing their scalability and robustness in real-world deployments. In contrast, AOPAD circumvents these issues by leveraging autoencoder models trained exclusively on bona fide data, in conjunction with one-class classifiers. By eliminating the reliance on attack data altogether, AOPAD offers a more principled and practical alternative, demonstrating enhanced resilience and generalization capabilities in realistic operating conditions.

III. Proposed method

We introduce AOPAD (Autoencoders for Open-Set Presentation Attack Detection) for real-world anti-spoofing, inspired by MORA [5], which uses autoencoders for gesture recognition. Our approach trains a single autoencoder on bona fide data, then clusters the extracted features and trains a one-class classifier for each cluster. During testing, new samples are classified as bona fide if they match any trained classifier, as shown in Figure 1. The following subsections detail each step of the approach.

A. Model Architecture

To clarify the structure of our pipeline, this section aims at detailing the architecture of the proposed model. It employs an autoencoder that takes images as input. The encoder is built upon MobileVIT [35], a lightweight adaptation of the Vision Transformer (VIT) [14], chosen for its low parameter count, which is an advantage given that training is performed exclusively on bona fide samples. The decoder comprises a sequence of 2D convolutional layers designed to reconstruct spatially meaningful features. The complete architecture is depicted in Figure 2.

Unlike conventional autoencoders reconstructing inputs, our model aims to reconstruct pertinent features crucial for face anti-spoofing, notably textures and reflectance. This can be mathematically represented as the process of learning a function $f\colon X \rightarrow Y$, where X denotes the input space and Y represents the feature space capturing relevant information for anti-spoofing tasks. Given the inherent disparity between inputs

and outputs, our model learns a mapping function. As the outputs encapsulate essential features for Face Anti-Spoofing (FAS), their activations encode valuable information about these features. During training, mean squared error (MSE) loss function $L_{MSE} = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2$ is employed to guide weight convergence, where Y_i represents the ground truth features for a given sample and $\hat{Y}_i$ denotes its reconstructed (predicted) features.

B. Feature Learning and Extraction

To extract relevant information, we use partial activations from an intermediate layer of the trained autoencoder, which capture spatial features important for the FAS task. After analyzing various layers in the decoding part, we found that intermediate layer 10 (shown in purple in Figure 2) provided the best performance. The dimensionality of extracted features varies by layer. During training, features are extracted only from bona fide data, while during testing, features are extracted from all samples for classification.

C. Mapping inputs into outputs

To effectively harness the capabilities of the trained autoencoder for feature extraction across diverse tasks beyond face anti-spoofing (FAS), our methodology is designed to be adaptable and extensible. The core principle of our approach revolves around the fundamental objective of transforming input data into feature maps teeming with task-specific information for a given application. This adaptable framework emphasizes our dedication to encapsulate key attributes relevant for any specific task. For the present work, we had a particular emphasis on texture and reflectance, commonly approached by FAS methods [50].

Our model encodes input data into feature maps, ensuring that activations from intermediate layers retain spatial information relevant to the task. This versatility allows easy extension to different domains by adjusting desired features and output representations. Mathematically, this adaptability can be represented succinctly:

Let I denote the input data, $F(I)$ represent the feature map extracted by the autoencoder, and O signify the set of desired output features tailored to the specific task. The mapping function F encodes relevant characteristics into the feature map, emphasizing task-specific attributes (A), as in Equation 1.

$$F(I) = A(I) \oplus O \tag{1}$$

Here, $\oplus$ symbolizes the concatenation operation. By providing a distinct set of output features O, our methodology seamlessly adapts to diverse tasks, effectively capturing essential information pertinent to each application.

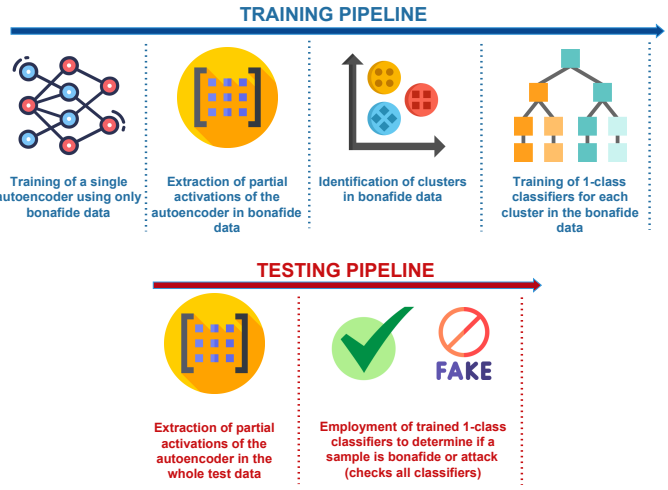


Fig. 1. AOPAD training and test phases. Training phase encompasses steps that range from the training of the single autoencoder for feature extraction to the training of 1-class classifiers for each cluster. Test phase only comprises the feature extraction and classification using all trained 1-class classifiers.

D. Face Anti-spoofing Features

The AOPAD method uses an autoencoder network to integrate key features for face antispoofing, focusing on extracting texture and reflectance information. The model is trained to predict these pivotal features, essential for building robust antispoofing mechanisms. Notable features include:

Binarized Statistical Image Features (BSIF): BSIF is a descriptor used to capture image texture by creating a binary code for each pixel. This is done by projecting local image patches onto a subspace, with basis vectors learned from natural images through independent component analysis, followed by binarization via thresholding [27]. BSIF has been successfully applied in object recognition and texture classification. Its performance depends on the region size, the number of binary patterns per neighborhood, and the rotation-invariant LBP operator size. For example, "BSIF $(3 \times 3 \times 6)$" refers to a 3x3 neighborhood, three binary patterns, and a size 6 LBP operator.

Single Scale Retinex (SSR): is an image enhancement technique used to adjust the dynamic range of an image, improving its appearance, brightness, contrast, sharpness, and visual quality [13]. It is commonly employed in computer vision and image processing applications. The algorithm is defined as the difference between the image at a given pixel and the center-surround average of that pixel. The final SSR image combines illumination and reflectance components, the last of which is obtained by dividing the original image by the estimated illumination component. This step helps to remove illumination effects from the image, enhancing details and textures and emphasizing the differences between real faces and synthetic ones. Each pixel of the final SSR image is produced by Equation 2.

$$SSR(x,y) = \log(I(x,y)) - \log(I(x,y) * G(x,y)) \tag{2}$$

where $I(x,y)$ represents the input image and $G(x,y)$ denotes a Gaussian filter.

Local Binary Patterns (LBP): is a texture descriptor that captures textures based on the local structures of digital images. To accomplish that, a window is examined around each pixel, and its value is compared with its neighbors, leading to the

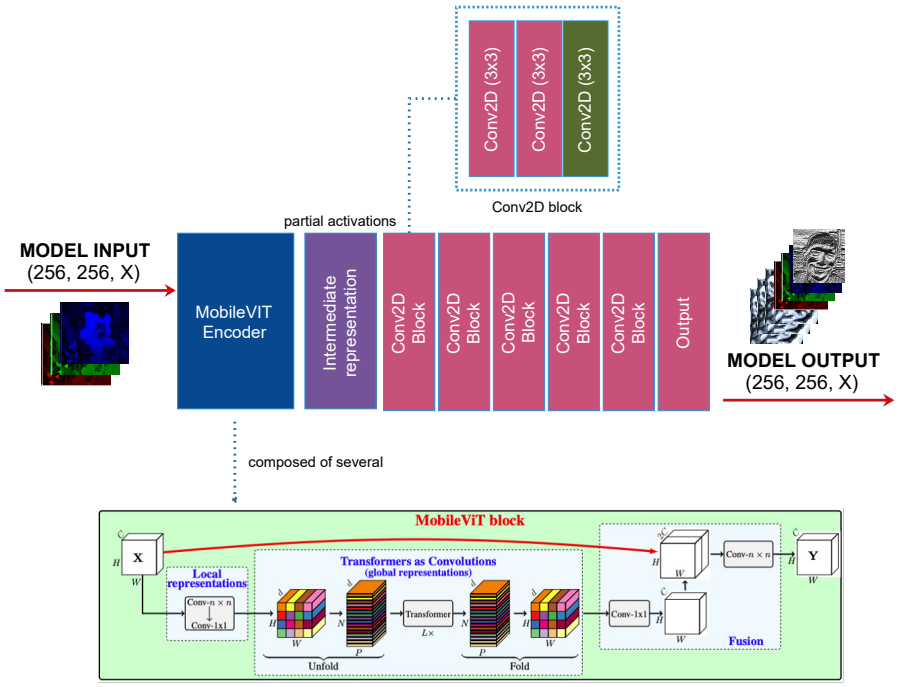


Fig. 2. AOPAD autoencoder architecture. The number of input and output channels depends on the data that is fed to the model.

generation of a binary code for each pixel. LBP has proven effective in various applications such as face recognition, texture classification, and object detection [40]. Parameters such as the window size and the number of neighbors regulate the code generation mechanism.

Short-wave Infrared (SWIR): is an imaging capture technology that captures light in the wavelength range of approximately 900 to 1700 nanometers [45]. SWIR images are captured by sensors sensitive to near-infrared and short-wave infrared regions of the electromagnetic spectrum, and SWIR wavelengths can penetrate through certain materials, such as fog, smoke, and even some types of clothing [45]. Additionally, various materials reflect SWIR light differently, enabling the detection of hidden objects or materials with distinct SWIR signatures. SWIR images provide valuable information beyond what visible light cameras capture, making them a useful tool in various applications across industries.

Figure 3 presents a frame from the HQ-WMCA dataset and the corresponding feature responses.

E. Clustering Activations

In an ideal scenario, during the training phase, features derived from partial activations of the trained autoencoder would be employed to train a single classification model, specifically designed as a one-class classifier. This classifier's objective would be to construct a hypersphere that effectively encapsulates the features exclusively from the bona fide samples.

The suitability of training a single one-class classifier is evident when the bona fide class information cohesively forms a singular group. However, the bona fide class may be fragmented in real-world scenarios, existing as separate clusters. Consequently, relying on a single hypersphere can compromise the method's performance. In order to encompass all bona fide data, an exceedingly large hypersphere would be generated, increasing the risk of inadvertently including attack data within its bounds, as depicted in Figure 4.

Hence, the proposed method adopts a strategic approach utilizing multiple one-class classifiers to address this concern. Initially, Principal Component Analysis (PCA) is utilized to reduce the dimensionality of the features. Let X denote the partial activations extracted using the autoencoder model for numerous samples, with dimensions $m \times n$, where m represents the sample count and n signifies the dimensionality of the activations. Following PCA application, we acquire the reduced feature matrix X_{PCA}. For each sample, the data is referred as X_{PCA_i}, as in Equation 3, where p represents the number of components.

$$X_{PCA_i} = [x_1, x_2, \ldots, x_p] \tag{3}$$

Subsequently, K-Means clustering is applied on X_{PCA}, with K determined by the silhouette score method [44]. This leads to $C = [c_1, c_2, \ldots, c_K]$ as a set of cluster centroids and $y = [y_1, y_2, \ldots, y_m]$ as cluster assignments. Then, each sample x_i is assigned to a specific cluster, generating the prediction y_i based on the nearest cluster centroid. After that, a one-class classifier is trained for each cluster. It is worth noting that this clustering process is solely implemented during training. All trained one-class classifiers are utilized for testing purposes to verify if the features fall within their respective hyperspheres. This process was previously depicted in Figure 1.

Fig. 3. A sample from HQ-WMCA and the respective feature responses. From left to right: grayscale, BSIF, SSR, LBP and SWIR

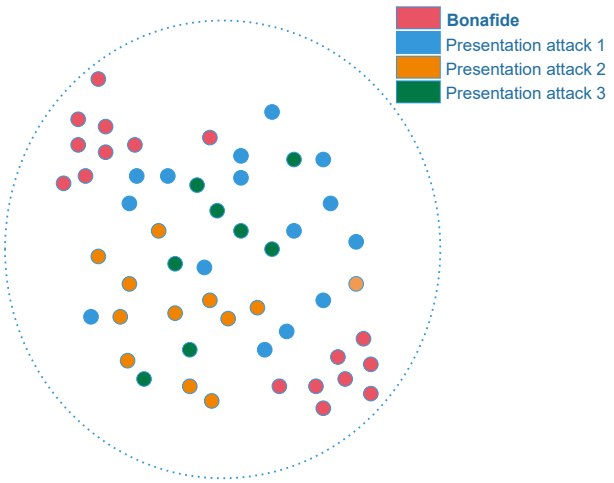


Fig. 4. In real-world situations, bona fide features could be dispersed into disjoint clusters. A single hypersphere may not encompass all of the bona fide data without encompassing attack samples.

F. One-class Classification

In the current approach, the primary goal of one-class classification is to distinguish genuine samples from potential attacks. Despite a previous step to reduce data dimensionality, the 1-class classification is executed on the original data, encompassing all its dimensions. As previously mentioned, all trained 1-class classifiers are utilized to assess whether a sample is classified as bona fide. It is important to note that a dedicated 1-class classifier is trained for each identified cluster.

We explored two methods for 1-class classification: (1) isolation forests and (2) one-class SVMs, with Isolation forests demonstrating superior performance and being used as 1-class classifiers for our method. Therefore, the classification step, according to our framework, follows Equation 4, where $\{OC_1, OC_2, \ldots, OC_K\}$ denotes a set of trained classifiers, K is the total number of classifiers (or clusters), x represents the partial activations of the autoencoder model and $\hat{y}$ represents the final output for a given sample.

$$\hat{y} = \begin{cases} bonafide & \text{if} \quad \text{OC}_i(x) = 1 \quad \text{for any} \quad i = 1, 2, \ldots, K \\ Attack & \text{otherwise} \end{cases} \tag{4}$$

It is important to mention that the behavior of the 1-class-classifiers is governed by the contamination parameter, determined through the computation of the silhouette score. Further details on tuning this hyperparameter can be found in section IV.

IV. Experiments and Results

This section provides a concise yet comprehensive overview of the evaluation process and results of the AOPAD method. Each subsection focuses on elucidating specific aspects, such as datasets, features, architecture inputs/outputs, hyperparameters, evaluation metrics, and results. Through structured presentation, the section facilitates a clear understanding of AOPAD's methodology and outcomes.

A. The datasets

The present approach was evaluated with six face antispoofing datasets. This evaluation followed protocols specified in the dataset guidelines.

RECOD Mobile Presentation Attack Dataset (RECOD-MPAD) [2] is a comprehensive FAS image dataset designed to simulate mobile device unlocking scenarios. It includes over 76,000 facial images from 24 subjects, with varied illumination and both indoor and outdoor environments. The dataset features bona fide images along with two attack types: print and screen attacks, captured using two different devices. This diversity in subjects, conditions, and attack types makes RECOD-MPAD a challenging benchmark for evaluating facial authentication systems.

Wide Multi-Channel Presentation Attack Dataset (WMCA) [16] is a FAS database containing 1,941 short video recordings from 72 subjects. In addition to RGB frames, it offers depth, infrared, and thermal data, enabling multi-modal research in face antispoofing. WMCA captures videos under diverse environmental and lighting conditions, making it valuable for evaluating the robustness of FAS algorithms. The dataset includes bona fide videos and nine attack types, such as fake heads, glasses, print, replay, rigid mask, flexible mask, paper mask, and wig.

High-Quality Wide Multi-Channel Presentation Attack Dataset (HQ-WMCA) [22] consists of 2,904 high-resolution video recordings from 51 subjects. Like WMCA, it includes multiple data channels, as RGB, depth, thermal, infrared, and short-wave infrared (SWIR) images, captured under varied lighting and environmental conditions. HQ-WMCA enriches the WMCA with new videos, offering a broader range of samples for analysis. In addition to the bona fide class, it

includes ten attack types: glasses, mannequins, prints, replay, rigid, flexible, paper, wigs, tattoos, and makeup.

OULU-NPU [7] is a mobile face presentation attack dataset specifically designed to reflect real-world usage scenarios. It includes videos of 55 subjects captured with six different mobile devices under varying illumination, background, and attack types (print and replay). It is partitioned into four protocols emphasizing generalization to unseen conditions, making it a valuable benchmark for evaluating open-set anti-spoofing methods.

MSU Mobile Face Spoofing Database (MSU-MFSD) [47] is a dataset used for face spoofing detection, consisting of 280 videos from 35 subjects. The videos cover both real access and two spoofing scenarios (print and video replay), captured with webcams and smartphones under various lighting conditions. Its simplicity makes it suitable for assessing methods' performance in constrained environments.

Replay-Attack [12] is a video dataset that includes 1,300 videos from 50 subjects, covering both controlled and adverse lighting conditions. It features three attack types (print, digital photo, and video replay) performed under hand-held and fixed-display configurations. Due to its standardized acquisition protocols and extensive use in prior work, it is widely adopted for benchmarking FAS systems.

Figure 5 illustrates representative samples from the datasets employed in this study. Although WMCA and HQ-WMCA offer multiple sensing channels, only a subset of these modalities was utilized in our experiments. The datasets consist of either image samples or short video sequences. Because these videos contain very few frames and exhibit minimal temporal variation, we extracted only the first frame from each sequence and applied the architecture shown in Figure 2 without modifying the model to process temporal information. On average, the evaluated videos include 60–64 frames (except for RECOD, which is an image database), roughly two seconds of recording, which results in negligible frame-to-frame differences. Indeed, most HQ-WMCA and WMCA samples display little to no observable changes across frames.

B. Inputs and outputs for each dataset

To assess the efficacy of AOPAD across various datasets, adjustments to the dimensions of both input and output of the architecture depicted in Figure 2 were imperative. Table I presents the set of inputs and the dimensionality of each input modality. In addition, it is important to notice that the selection of model outputs aimed at integrating pertinent information essential for the face antispoofing task, particularly concerning textures and reflectance. One could notice that SWIR channels are used as outputs for the model trained in WMCA despite the dataset not providing this information. SWIR information from the HQ-WMCA dataset was used to train AOPAD in this dataset. So, only the videos with correspondence between the two datasets were employed in this test, resulting in 1326 from the 1941 original WMCA videos. Despite using SWIR information from HQ-WMCA, the protocol of WMCA was still respected, with the same separation between attack and bona fide videos.

C. Determining Hyperparameters

With AOPAD, our focus was on replicating real-world scenarios without prior knowledge of any specific presentation attack. Therefore, our methodology avoids utilizing validation data derived from attacks during the training phase. To determine model hyperparameters, we adopted a simple yet effective approach, considering the minimization of reconstruction error for validation purposes.

When determining the hyperparameters for AOPAD, we also take into account the number of data clusters, which represents how the genuine data is distributed, and the contamination parameter used in the one-class classifiers. The contamination parameter helps to determine the threshold for classifying data points as outliers or inliers, reflecting the expected proportion of outliers in the training dataset. The process to determine these two hyperparameters was based on the activations extracted with an intermediate layer of the autoencoder model.

The silhouette score is a commonly used metric to evaluate the quality of clustering. It assesses how well each data point fits within its assigned cluster compared to other clusters. A higher silhouette score indicates that the data point is well-matched to its cluster and poorly matched to neighboring clusters. This measure provides insights into the compactness and separability of the clusters. By employing the silhouette score, we can determine the optimal hyperparameters for our approach, such as the number of clusters and the contamination of classifiers. This metric is computed by Equation 5,

$$\textit{silhouette score} = \frac{(b-a)}{max(a,b)} \tag{5}$$

Where a represents the intra-cluster and b represents the nearest-cluster mean distances, respectively.

First, we utilized the silhouette score to determine the optimal number of clusters in the data. By training a K-means algorithm for each possible number of clusters, ranging from 1 to 15, we could compute the silhouette score for each point and calculate the average score across all potential numbers of clusters. The number of clusters that yielded the highest average silhouette score was considered the optimal choice. This method allows us to objectively assess and select the number of clusters that provide the best clustering performance for the data. The pseudocode corresponding to this procedure is provided in Algorithm 1.

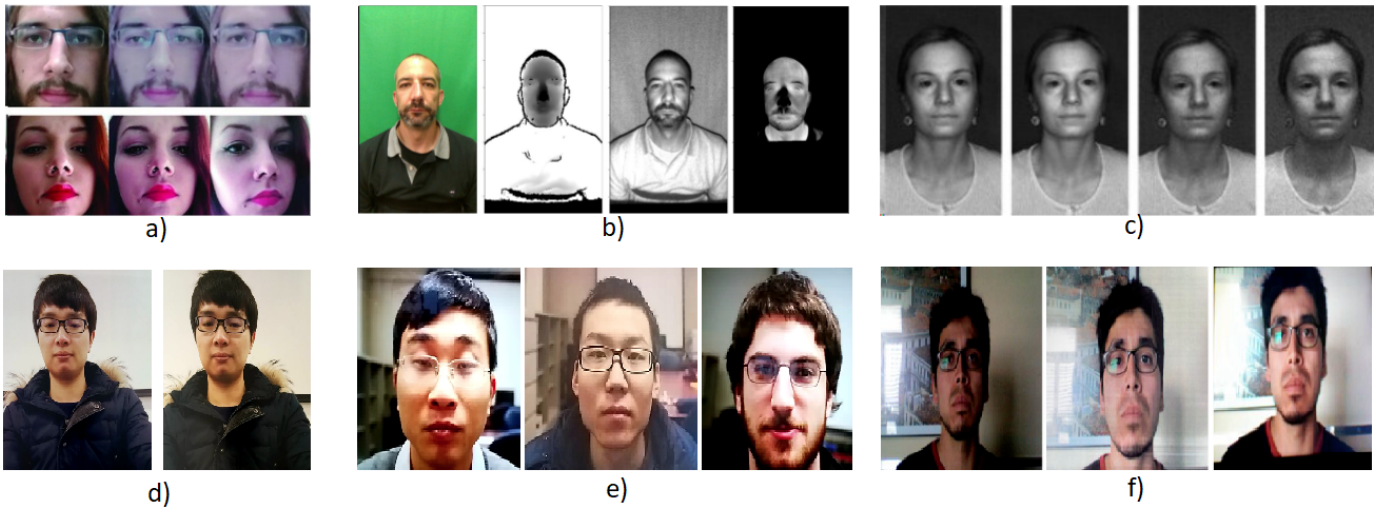


Fig. 5. Samples from the employed datasets. a) Face images from RECOD-MPAD. b) Multichannel inputs from WMCA (RGB, depth, infrared, thermal). c) SWIR channels from HQ-WMCA. d) Print attack samples from OULU-NPU. e) Different subjects on MSU-MFSD. f) Illumination variation on Replay-Attack

TABLE I
AOPAD INPUTS AND OUTPUTS FOR THE EVALUATED DATASETS

Dataset	Inputs	Input Dimension	Output	Output Dimension
RECOD-MPAD	Red, Green, Blue	$(256 \times 256 \times 3)$	BSIF, SSR and LBP	$(256 \times 256 \times 7)$
HQ-WMCA	Grayscale, SWIR	$(256 \times 256 \times 43)$	Grayscale, SWIR, BSIF, SSR, LBP	$(256 \times 256 \times 48)$
WMCA	Red, Green, Blue	$(256 \times 256 \times 3)$	Grayscale, SWIR	$(256 \times 256 \times 43)$
OULU-NPU	Red, Green, Blue	$(256 \times 256 \times 3)$	BSIF, SSR and LBP	$(256 \times 256 \times 7)$
MSU-MFSD	Red, Green, Blue	$(256 \times 256 \times 3)$	BSIF, SSR and LBP	$(256 \times 256 \times 7)$
Replay-Attack	Red, Green, Blue	$(256 \times 256 \times 3)$	BSIF, SSR and LBP	$(256 \times 256 \times 7)$

Algorithm 1 Selecting the optimal number of clusters using silhouette score

Ensure: Optimal number of clusters K^*
  $K_{\max} = 15$
  for $k = 1$ to $K_{\max}$ **do**
    Train K-means on X with k clusters
    Compute silhouette score s_i for each point $x_i \in X$
    Compute average silhouette score $\bar{s}_k = \frac{1}{|X|} \sum_i s_i$
  end for
  $K^* \leftarrow \arg \max_{k \in \{1,\ldots,K_{\max}\}} \bar{s}_k$
  return K^*

With the number of clusters X determined, we trained X isolation forest classifiers using a range of contamination values from 0.01 to 0.3, thereby covering a broad spectrum of possible anomaly thresholds. This interval allows us to explore realistic anomaly proportions without assuming an unrealistically high or negligible anomaly rate. For each contamination value, some data points are classified as inliers and others as outliers. We then evaluated the quality of the inlier points using again the silhouette score, which measures how well data points fit within their clusters, balancing both cohesion and separation. The contamination value that maximized the silhouette score was selected as the optimal parameter. In cases where multiple values yielded the same score, we chose the lowest contamination value, based on the assumption that it results in fewer data points being marked as anomalies, thus forming larger and potentially more meaningful clusters of inliers. Importantly, we observed that the method produced consistent results across different contamination values, reinforcing that the parameter selection was not arbitrary. The pseudocode corresponding to this procedure is provided in Algorithm 2.

Algorithm 2 Selecting the optimal contamination value using silhouette score

Ensure: Optimal contamination value c^*
  Define the contamination set $\mathcal{C} = \{0.01, 0.02, \ldots, 0.30\}$
  for each contamination value $c \in \mathcal{C}$ **do**
    Train an Isolation Forest with contamination c
    Classify each data point as inlier or outlier
    Extract the set of inliers I_c
    Compute silhouette score s_c on inlier points I_c
  end for
  Identify the maximum silhouette score $s_{\max} = \max_{c \in \mathcal{C}} s_c$
  Let $\mathcal{C}_{\max} = \{ c \in \mathcal{C} \mid s_c = s_{\max} \}$ {contamination values tied for best score}
  Select $c^* = \min(\mathcal{C}_{\max})$ {choose lowest contamination value in case of ties}
  return c^*

Table II presents the number of clusters and the chosen contamination value for all evaluated datasets.

To ensure that the hyperparameter selection process remained unbiased, silhouette scores were computed strictly on bona fide training data, without incorporating any attack

TABLE II
AOPAD NUMBER OF CLUSTERS AND CONTAMINATION FOR THE EVALUATED DATASETS

Dataset	Number of Clusters	Contamination
RECOD-MPAD	2	0.14
HQ-WMCA	2	0.19
WMCA	1	0.21
OULU-NPU	4	0.18
MSU-MFSD	2	0.06
Replay-Attack	2	0.08

samples at any stage of cluster evaluation or parameter search. By selecting the contamination value that maximized the silhouette score of the inlier set, we obtained the parameter configuration that best preserved the intrinsic structure of the bona fide data.

D. Evaluation metrics

To properly evaluate AOPAD and compare our method with literature methods, we gathered several metrics commonly reported by anti-spoofing approaches. Table III presents these metrics, along with the equation to compute each of them. To compute these metrics, the bona fide samples represent the positive class. The protocol to compute Equal Error Rate (EER) and HTER is the one proposed in [19].

E. Results

This section presents the results from our AOPAD experiments, highlighting its effectiveness across various datasets. While some results may not be directly comparable to literature methods due to different evaluation protocols, they demonstrate the applicability of our approach. It is noteworthy that within the realm of anti-spoofing methodologies, most literature methods adhere to the rigorous Leave-One-Attack-Type-Out (LOO) methodology, involving the systematic removal of one attack instance per experiment. In contrast, our methodology predominantly involved partitioning the data into bona fide and attack categories, where the latter encompassed all forms of presentation attack types. However, to ensure fairness in the comparison with established literature methodologies, we conducted an extra experiment applying the same LOO protocol for HQ-WMCA. The insights and results from these experiments are discussed in the following sections.

1) ***Experiments on RECOD-MPAD****:* The performance of AOPAD on RECOD-MPAD [2] is shown in Table IV, alongside other methods from the literature. While all these methods outperform AOPAD, they are supervised approaches that struggle with unseen presentation attacks. Despite being the least accurate on RECOD-MPAD, AOPAD still achieved low EER and HTER values, demonstrating good performance without any attack examples during training. Since some methods may not report all metrics, comparisons should consider those reported by both AOPAD and the other approaches for fairness.

The RECOD-MPAD dataset experiment unveiled the remarkable resilience of AOPAD against unseen presentation attacks, as evidenced by its performance in Equal Error Rate (EER) and Half Total Error Rate (HTER). Noteworthy is AOPAD's attainment of an EER of 0.034 and an HTER of 0.041, achieved without exposure to attack examples during training. This underscores AOPAD's promising efficacy in real-world applications. It is imperative to conduct fair comparisons by considering all reported metrics and employing a uniform evaluation protocol to ensure a thorough assessment of AOPAD's performance against other methods.

2) ***Experiments on WMCA****:* The performance of AOPAD for WMCA [16] dataset can be seen in Table V, alongside literature methods. The literature results in Table V follow a LOO protocol, leaving one attack type out of the training and checking if the method can identify it. Despite still being an advantage over the AOPAD protocol, which uses only bona fide information, the comparison is fairer than what is reported in RECOD-MPAD. One can notice that for this dataset, AOPAD was only surpassed by MCCNN-GMM [17] approach in terms of ACER metric.

The assessment of AOPAD's performance on the WMCA dataset provides valuable insights. Unlike AOPAD's protocol, which exclusively relies on genuine information, the literature methods get some advantage, since they adopt a Leave-One-Attack-Type-Out (LOO) protocol. Despite this disparity, AOPAD demonstrates commendable performance. This performance indicates that our strategy to map RGB inputs into SWIR frames led the model activations to be imbued with information that is relevant for spoofing attacks.

3) ***Experiments on HQ-WMCA****:* The performance of AOPAD for HQ-WMCA [22] dataset can be seen in Table VI, alongside other methods from the literature. All the results presented in this table, except for AOPAD, are related to methods trained not to handle unseen attacks. Despite presenting the highest ACER score among them, AOPAD obtains good performance, rivaling methods that are trained with attack labels and unsuitable for handling real-world situations in which attacks are unknown.

This experiment, much like its counterpart on the WMCA dataset, illuminates the wealth of valuable information accessible through SWIR channels. Furthermore, it underscores the exceptional capability of our method to rival established literature methods, despite operating within challenging constraints: training without any attack information and adeptly handling unseen presentation attack types.

4) ***Experiments on OULU-NPU****:* Table VII presents the performance of AOPAD on the OULU-NPU [7] dataset. As for previous datasets, the literature results were obtained using a Leave-One-Attack-Type-Out (LOO) protocol, where one attack type is excluded during training and used only for testing. This evaluation setup differs from ours and prevents a fully fair comparison. Nevertheless, AOPAD demonstrates competitive performance, being comparable with most of the existing methods despite the less favorable evaluation conditions.

5) ***Experiments on MSU-MFSD****:* Table VIII presents the performance of AOPAD on the MSU-MFSD dataset [47]. As most existing studies report cross-domain or supervised results for this dataset, we include supervised results from the literature to enable a fairer comparison. Although AOPAD is trained without any attack data, it achieves the second-lowest reported EER, being very close to the best-performing

TABLE III
METRICS USED TO EVALUATE AOPAD AND COMPARE PERFORMANCE WITH LITERATURE METHODS. FN, TP, FP, AND TN ARE FALSE NEGATIVES, TRUE POSITIVES, FALSE POSITIVES, AND TRUE NEGATIVES, RESPECTIVELY.

Metric	Equation
Average Classification Error Rate (ACER)	$\frac{(APCER+BPCER)}{2}$
Attack Presentation Classification Error Rate (APCER)	$\frac{FP}{(FP+TN)}$
Bona Fide Presentation Classification Error Rate (BPCER)	$\frac{FN}{(FN+TP)}$
Equal Error Rate (EER)	FRR when it equals FAR in dev set
HTER	$\frac{(FAR+FRR)}{2}$ using contamination determined to achieve EER

TABLE IV
AOPAD RESULTS FOR RECOD-MPAD DATASET. AOPAD RESULTS WERE COMPUTED USING THE CONTAMINATION PARAMETER OF 0.14, OBTAINED USING THE SILHOUETTE SCORE. *PT MEANS PRE-TRAINED

Approach	ACER	BPCER	APCER	HTER	EER
Color LBP [6]	-	-	-	0.028	0.027
PT* CNN [2]	-	-	-	0.0108	0.0094
PBCNN [2]	-	-	-	**0.0054**	**0.0037**
AOPAD	0.160	0.138	0.183	0.041	0.034

TABLE V
AOPAD RESULTS FOR WMCA DATASET. AOPAD RESULTS WERE COMPUTED USING THE CONTAMINATION PARAMETER OF 0.21, OBTAINED USING SILHOUETTE SCORE.

Approach	ACER	BPCER	APCER	HTER	EER
MC-LBP-SVM [17]	0.195	0.002	0.389	-	-
MCCNN(BCE) [16]	0.310	**0.000**	0.620	-	-
MCCNN-GMM [17]	**0.097**	0.039	0.154	-	-
AOPAD	0.185	0.280	**0.090**	0.085	0.096

TABLE VI
AOPAD RESULTS FOR HQ-WMCA DATASET. AOPAD RESULTS WERE COMPUTED USING THE CONTAMINATION PARAMETER OF 0.19, OBTAINED USING THE SILHOUETTE SCORE.

Approach	ACER	BPCER	APCER	HTER	EER
MA-NET [31]	0.068	0.026	0.068	-	-
DeepPixBiS [19]	0.060	0.037	0.082	-	-
Conv-MLP [46]	**0.014**	0.012	**0.016**	-	-
AR-MLP [20]	0.016	**0.009**	0.021	-	-
AOPAD	0.070	0.069	0.071	0.047	0.024

TABLE VII
AOPAD RESULTS FOR OULU-NPU DATASET. AOPAD RESULTS WERE COMPUTED USING THE CONTAMINATION PARAMETER OF 0.18, OBTAINED USING THE SILHOUETTE SCORE.

Approach	ACER	BPCER	APCER	HTER	EER
IQM-GMM [39]	0.350	0.166	0.534	-	-
Baweja et al [9]	0.354	0.106	0.602	-	-
Lim et al [29]	0.280	0.205	0.369	-	-
AAE [24]	0.331	0.401	0.264	-	-
OC-SCMNet [25]	**0.140**	**0.116**	**0.164**	-	-
AOPAD	0.264	0.282	0.246	0.245	0.228

method. This strong result may stem from the dataset's small size and the efficiency of our lightweight MobileVIT-based architecture, which performs well even with limited data. Moreover, the dataset's constrained nature, with few subjects and capture devices, may have encouraged more compact feature clusters, further enhancing performance. It is worth noting that, for MSU-MFSD, prior work reports only EER.

*6) **Experiments on Replay-Attack**:* Table IX shows the performance of AOPAD on the Replay-Attack dataset [12], alongside results from supervised methods. Although these methods are not particularly complex, AOPAD does not match the performance of the best reported approaches. This gap may be attributed to differences in the evaluation protocol: while AOPAD operates without access to any attack data, the compared methods are trained with full supervision, including attack samples. In addition, our methodology only extracts information from a single frame of the videos. Employing an architecture that extracts temporal information could enhance the performance of our method.

*7) **Going deeper on HQ-WMCA - Fair comparison with literature**:* Although AOPAD achieved precise results across various datasets, direct comparison with existing literature methods was hindered by variations in the employed protocols. To address this discrepancy, an additional test was conducted to replicate the Leave-One-Attack-Type-Out (LOO) protocol commonly seen in the literature. By aligning our experimental approach with established practices, we were able to make a more equitable assessment of our method alongside existing techniques. We aligned our methodology with existing literature standards by using the HQ-WMCA dataset with the Leave-One-Attack-Type-Out (LOO) protocol. Figure 6 illustrates the performance of various literature methods (MA-NET [31], DeepPixBis [19], Conv-MLP [46], AR-MLP [20]) alongside AOPAD. Adhering to this protocol not only ensured a fairer evaluation but also highlighted AOPAD's competitive performance on HQ-WMCA (in red), achieving lower ACER values compared to the other evaluated methods.

This experiment underscores the importance of harmonizing evaluation protocols to ensure fair comparisons across different methods. Under these standardized conditions, AOPAD demonstrated strong performance, achieving favorable ACER results relative to existing approaches. These findings reaffirm not only the effectiveness but also the significant potential of AOPAD in real-world biometric security scenarios.

F. Ablation Study

In the development of the AOPAD approach, two different backbone networks were meticulously tested for our autoencoder. Furthermore, we assessed the effectiveness of

TABLE VIII
AOPAD RESULTS FOR MSU-MFSD DATASET. AOPAD RESULTS WERE COMPUTED USING THE CONTAMINATION PARAMETER OF 0.06, OBTAINED USING THE SILHOUETTE SCORE.

Approach	ACER	BPCER	APCER	HTER	EER
EBDG [15]	-	-	-	0.095	-
LBP-TOP [37]	-	-	-	0.058	-
IADG [51]	-	-	-	0.054	-
NAS-FAS [49]	-	-	-	0.116	-
RGGS [26]	-	-	-	**0.023**	-
HiPTune [34]	-	-	-	-	0.065
AOPAD	0.025	0.029	0.021	0.025	0.024

TABLE IX
AOPAD RESULTS FOR REPLAY-ATTACK DATASET. AOPAD RESULTS WERE COMPUTED USING THE CONTAMINATION PARAMETER OF 0.08, OBTAINED USING THE SILHOUETTE SCORE.

Approach	ACER	BPCER	APCER	HTER	EER
LBP+YcbCr [8]	0.059	0.051	0.068	-	-
Fake-NET [1]	**0.007**	0.007	**0.008**	-	-
FGCN [11]	0.029	0.024	0.034	-	-
MCAN [11]	0.009	0.007	0.011	-	-
MCAN+FGCN [11]	0.010	**0.006**	0.013	-	-
AOPAD	0.044	0.058	0.030	0.038	0.034

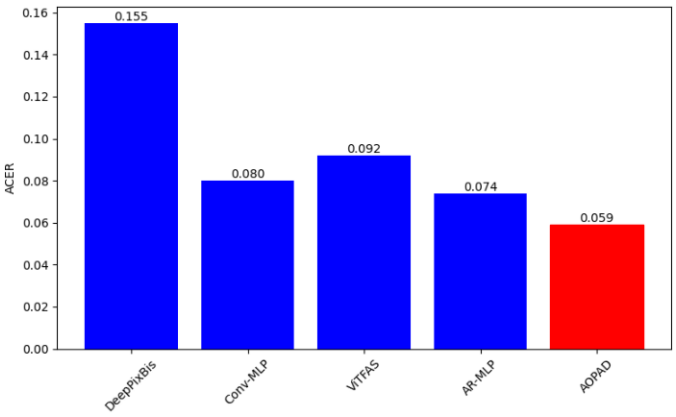


Fig. 6. AOPAD results for HQ-WMCA dataset following LOO (unseen attack) protocol. AOPAD results were computed using the contamination parameter of 0.17, obtained using silhouette score.

our method, including the clustering step. These experiments substantiated the rationale behind each stage of our method, ultimately enabling us to enhance the performance of AOPAD. Table X presents the ACER metric value for each variation of AOPAD. Bold results were the ones chosen to be employed for each dataset since they led the method to achieve better results.

In all experiments, except for the ones performed in HQ-WMCA, the incorporation of clustering techniques presented a substantial enhancement, showcasing a remarkable decrease in the ACER metric. The performance improvement due to cluster approach can be associated with the feature extraction process, that for all datasets, expect for HQ-WMCA, generated separate and distinct feature clusters due to illumination, device, subject and background variantions, as depicted in Figure 7 for the WMCA dataset. Regarding HQ-WMCA, since

TABLE X
ABLATION STUDY CONDUCTED IN SIX DIFFERENT DATASETS. THE COLUMNS REPRESENT THE EMPLOYMENT OF THE CLUSTERING METHOD OR NOT. IN ADDITION, IT EVALUATES THE EMPLOYED BACKBONE TO ENCODE DATA (VIT [14] OR MOBILEVIT [35])

Dataset	Clustering	Backbone	ACER
RECOD-MPAD	**Yes**	**MobileVIT**	**0.160**
	No	MobileVIT	0.208
	Yes	VIT	0.172
	No	VIT	0.212
WMCA	**Yes**	**MobileVIT**	**0.185**
	No	MobileVIT	0.220
	Yes	VIT	0.196
	No	VIT	0.236
HQ-WMCA	**Yes**	**MobileVIT**	**0.070**
	No	MobileVIT	0.073
OULU-NPU	**Yes**	**MobileVIT**	**0.264**
	No	MobileVIT	0.302
	Yes	VIT	0.290
	No	VIT	0.324
MSU-MFSD	**Yes**	**MobileVIT**	**0.025**
	No	MobileVIT	0.055
	Yes	VIT	0.036
	No	VIT	0.058
Replay-Attack	**Yes**	**MobileVIT**	**0.044**
	No	MobileVIT	0.048
	Yes	**VIT**	**0.044**
	No	VIT	0.049

the best silhouette score was obtained with a single cluster, this experiment appears to offer limited insights, and only the variation of the backbone was evaluated.

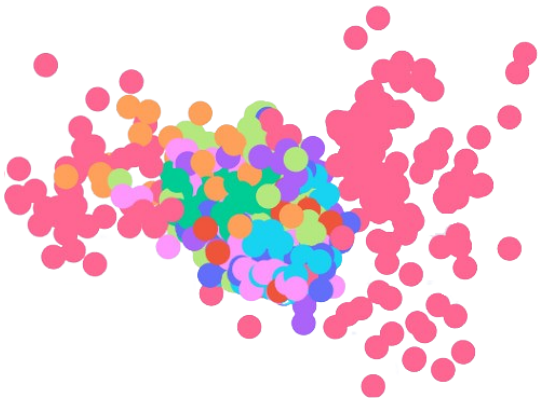


Fig. 7. Projection of WMCA Features, Dark pink spheres represent bona fide data, spread into two blocks. The other colors, mostly present in the center, are associated with presentation attack types.

It is important to notice that MobileVIT outperformed VIT across all tested datasets, except for Replay-Attack, where both models achieved the same ACER. The superior performance of MobileVIT in our setting is largely due to VIT's reliance on large-scale datasets to learn effective representations. VIT tends to underfit or overfit in data-scarce scenarios like ours, since the datasets are small and we only use bona fide information. In contrast, MobileVIT is less data dependent, being able to capture fine-grained patterns even with limited data. This makes it better suited for applications like ours, where data is scarce. To further support this observation, we conducted an additional experiment on HQ-WMCA, the dataset with

the largest number of video samples. In this analysis, we progressively increased the amount of bona fide training data by randomly sampling subsets ranging from 1 to 555 samples. For each subset size, we trained two models, one with VIT and another with MobileVIT as the backbone, and measured the resulting ACER. The outcomes, shown in Figure 8, reveal a clear trend: VIT exhibits a nearly linear reduction in ACER as more training samples are provided, yet it fails to reach the performance attained by MobileVIT. In contrast, MobileVIT displays a steep performance improvement even when trained with very small sample sizes, suggesting that it is more effective in data-constrained settings.

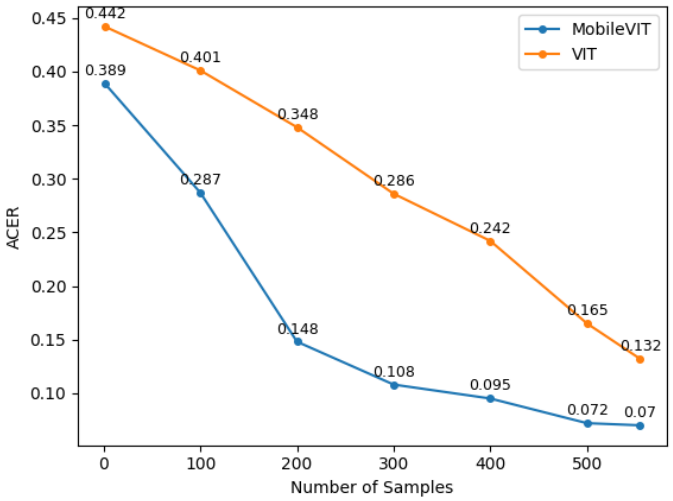


Fig. 8. ACER decay for a random incremental number of bonafide training samples. VIT and MobileVIT backbones are being evaluated.

G. AOPAD limitation: cross-dataset scenario

AOPAD demonstrated precision in delivering accurate results across all assessed datasets, effectively recognizing most inputs associated with unseen presentation attacks. While maintaining commendable performance, AOPAD encountered challenges in a cross-dataset context, a predictable outcome due to our method's dependence on autoencoder models for feature extraction.

These models are designed to minimize reconstruction errors relative to the input. However, in our cross-dataset evaluation, the model trained on RGB images from the WMCA dataset struggled to extract high-quality features when compared to the intra-dataset setting. This degradation is largely attributed to differences in acquisition conditions, including camera viewpoint, subject distance, angles, illumination, and image resolution. The trained model was tested on the RECOD-MPAD, OULU-NPU, MSU-MFSD, and Replay-Attack datasets. As expected, the drop in feature quality led to reduced performance across all datasets, as shown in Table XI, though AOPAD still achieved reasonable results on these tests.

Although the model shows slightly lower performance in cross-dataset evaluations, this does not undermine its applicability in real-world scenarios. In such experiments, each dataset introduces distinct viewpoints and sensor characteristics that can substantially affect the extracted feature representations. Autoencoders, in particular, are sensitive to these variations due to their reconstruction-based nature, which results in degraded performance when exposed to unfamiliar input distributions. It is important to emphasize, however, that this limitation is most relevant in academic cross-dataset protocols, where acquisition conditions are deliberately varied to assess generalization. Moreover, the degradation was especially pronounced in terms of BPCER, while APCER remained comparatively stable. This indicates that AOPAD maintains robustness against presentation attacks, **even across different attack types** present in the evaluated datasets, though at the cost of reduced capability to accurately recognize bona fide samples.

TABLE XI
AOPAD RESULTS FOR CROSS-DATASET EXPERIMENT. FOR EACH DATASET, WE REPORT THE INTRA-DATASET RESULT ALONG WITH THE CROSS-DATASET EXPERIMENT. ALL EVALUATION CONSIDERED A MODEL TRAINED ON WMCA DATASET.

Approach	ACER	BPCER	APCER
AOPAD-RECOD-MPAD	0.160	0.138	0.183
AOPAD-Cross	0.211	0.230	0.191
OULU-NPU	0.264	0.282	0.246
OULU-NPU-Cross	0.313	0.364	0.262
MSU-MFSD	0.025	0.029	0.021
MSU-MFSD-Cross	0.054	0.084	0.024
Replay-Attack	0.044	0.058	0.030
Replay-Attack-Cross	0.072	0.104	0.040

H. Facing adversarial attacks

To assess the robustness of AOPAD against adversarial perturbations, we performed an additional experiment on the RECOD-MPAD dataset. In this evaluation, adversarial noise was introduced into both test attack and test bona fide samples using the Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD). For FGSM, the perturbation intensity was regulated by the epsilon parameter, which was varied from 0.00 (no perturbation) to 0.20 in steps of 0.05. In the case of PGD, the perturbation intensity was also controlled by epsilon, varied over the same range, while the number of iterations was adjusted between 5 and 30 to analyze its impact. The epsilon parameter defines a normalized [0,1] bound that limits how far each pixel, also scaled to the range [0, 1], can be perturbed from its original value.

Although AOPAD relies on autoencoders and one-class classification, approaches typically sensitive to input perturbations, it exhibited notable resilience to adversarial noise up to a certain threshold for both attack methods. In particular, the model sustained reasonable performance for epsilon values up to 0.10, even under a high number of iterations in the case of PGD (30). Beyond these thresholds, performance deteriorated sharply, primarily due to a substantial increase in BPCER caused by the distortions introduced into the input images, as reported in Tables XII and XIII. Interestingly, the comparatively lower degradation in APCER indicates that adversarial perturbations did not increase the model's susceptibility to

attacks but rather impaired its ability to correctly recognize bona fide inputs. This suggests that AOPAD possesses a degree of robustness against adversarial interference.

TABLE XII
AOPAD RESULTS FOR ADVERSARIAL ATTACKS USING FGSM IN RECOD-MPAD.

epsilon	ACER	BPCER	APCER
0.00	0.160	0.138	0.183
0.05	0.182	0.168	0.196
0.10	0.256	0.298	0.218
0.15	0.388	0.554	0.222
0.20	0.467	0.686	0.248

TABLE XIII
AOPAD RESULTS FOR ADVERSARIAL ATTACKS USING PGD IN RECOD-MPAD.

epsilon	epochs	ACER	BPCER	APCER
0.00	**5**	0.160	0.138	0.183
0.05	**5**	0.180	0.172	0.188
0.10	**5**	0.248	0.292	0.204
0.15	**5**	0.315	0.394	0.236
0.20	**5**	0.370	0.487	0.252
0.00	**15**	0.160	0.138	0.183
0.05	**15**	0.198	0.206	0.190
0.10	**15**	0.231	0.248	0.214
0.15	**15**	0.346	0.447	0.244
0.20	**15**	0.434	0.609	0.258
0.00	**30**	0.160	0.138	0.183
0.05	**30**	0.249	0.295	0.202
0.10	**30**	0.319	0.402	0.235
0.15	**30**	0.416	0.564	0.268
0.20	**30**	0.530	0.782	0.277

Figure 9 illustrates an example from the RECOD-MPAD dataset subjected to PGD perturbations under different thresholds. As the epsilon parameter increases, the perturbations introduce progressively more noticeable artifacts into the images.

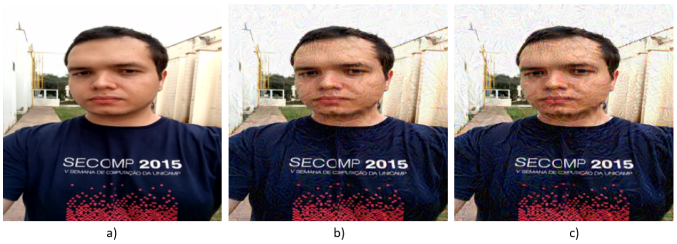


Fig. 9. Image deterioration due to PGD artifact generation for 30 epochs. a) Original image. b) epsilon 0.1. c) epsilon 0.2.

V. CONCLUSIONS AND FUTURE WORK

This paper introduces Autoencoders for Open-Set Presentation Attack Detection (AOPAD), a novel approach tailored to address anti-spoofing challenges prevalent in real-world scenarios. AOPAD stands out for its ability to handle unseen attack types and its independence from attack data during training, making it particularly suitable for scenarios where such data is scarce or unknown. Inspired by MORA's successful application of autoencoder networks in recognizing video gestures, AOPAD utilizes bona fide data exclusively during training to construct a robust anti-spoofing mechanism. By training a single autoencoder model on bona fide data and subsequently organizing its features into clusters, AOPAD offers a promising solution for identifying presentation attacks without the need for prior knowledge of specific attack types.

Evaluation across diverse datasets, including RECOD-MPAD, WMCA, HQ-WMCA, OULU-NPU, MSU-MFSD and Replay-Attack, validates AOPAD's efficacy in delivering accurate results comparable to supervised methods, even when faced with unseen attack types. AOPAD consistently matches or surpasses existing methods in metrics such as ACER, highlighting its potential for real-world deployment.

Autoencoder-based methods, including AOPAD, are more sensitive to domain shifts such as strong illumination, viewpoint, and resolution changes, as well as sensor characteristics. Cross-dataset experiments reveal some decline in performance, underscoring the need to improve adaptability and robustness when applying the model to unseen datasets. To address this, several promising strategies can be employed: (i) domain adaptation techniques to align latent feature distributions between source and target datasets, reducing dataset-specific biases; (ii) data augmentation that simulates cross-domain variability, such as changes in lighting, resolution, and viewpoint; and (iii) latent feature normalization or projection into a shared hyperspace to mitigate sensitivity to dataset-specific characteristics. Systematic exploration of these strategies in future work could strengthen AOPAD's cross-dataset generalization, making it even more resilient and reliable in diverse real-world scenarios.

In conclusion, AOPAD offers extensive potential applications in real-world scenarios, particularly in fortifying the security of facial recognition systems across various sectors. Its ability to detect presentation attacks without specific attack type knowledge makes it valuable for safeguarding user accounts and sensitive information in digital environments, being applicable in tasks such as mobile banking, e-commerce, and social media platforms. Additionally, AOPAD could enhance access control systems in high-security environments, reinforcing perimeter security and protecting critical assets. By leveraging AOPAD's robust anti-spoofing capabilities, organizations can advance the adoption of facial recognition technology, bolster security measures, and preserve user privacy across diverse contexts.

VI. ACKNOWLEDGMENTS

We thank the financial support of the São Paulo Research Foundation (Fapesp) under the Horus Grant #2023/12865-8.

REFERENCES

[1] M. Alshaikhli, O. Elharrouss, S. Al-Maadeed and A. Bouridane, "Face-Fake-Net: The Deep Learning Method for Image Face Anti-Spoofing Detection : Paper ID 45," 2021 9th European Workshop on Visual Information Processing (EUVIP)

[2] W. Almeida, F. Andaló, R. Padilha, G. Bertocco, W. Dias, R. Torres, J. Wainer, A. Rocha, "Detecting face presentation attacks in mobile devices with a patch-based CNN and a sensor-aware loss function", in PLOS ONE, vol. 15, pp. 1-24, 2020, doi: doi.org/10.1371/journal.pone.0238058

[3] S. R. Arashloo, J. Kittler and W. Christmas, "An Anomaly Detection Approach to Face Spoofing Detection: A New Formulation and Evaluation Protocol," in IEEE Access, vol. 5, pp. 13868-13882, 2017, doi: 10.1109/ACCESS.2017.2729161.

[4] S. Rahimzadeh Arashloo, "Unknown Face Presentation Attack Detection via Localized Learning of Multiple Kernels," in IEEE Transactions on Information Forensics and Security, vol. 18, pp. 1421-1432, 2023, doi: 10.1109/TIFS.2023.3240841.

[5] I. L. O. Bastos, V. H. C. Melo, G. R. Gonçalves and W. Robson Schwartz, "MORA: A Generative Approach to Extract Spatiotemporal Information Applied to Gesture Recognition," 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Auckland, New Zealand, 2018, pp. 1-6, doi: 10.1109/AVSS.2018.8639143.

[6] Boulkenafet Z, Komulainen J, Lei L, Xiaoyi F, Hadid A, "Face anti-spoofing based on color texture analysis". In: IEEE Int. Conf. Image Process.; 2015. p. 2636–2640.

[7] Boulkenafet Z, Komulainen J, Hadid A, "OULU-NPU: A Mobile Face Presentation Attack Database with Real-World Variations". In: International Conference on Automatic Face Gesture Recognition; 2017. p. 612–618.

[8] Z. Boulkenafet, Komulainen J, Z. Akhtar, A. Benlamoudi, D. Samai, S. E. Bekhouche, "A competition on generalized software-based face presentation attack detection in mobile scenarios," 2017 IEEE International Joint Conference on Biometrics (IJCB), Denver, CO, USA, 2017, pp. 688-696.

[9] Baweja, P. Oza, P. Perera, and V. M. Patel, "Anomaly detection-based unknown face presentation attack detection," IEEE International Joint Conference on Biometrics (IJCB); 2020.

[10] H. Chen, G. Hu, Z. Lei, Y. Chen, N. M. Robertson, and S. Z. Li, "Attention-based two-stream convolutional networks for face spoofing detection," IEEE Trans. Inf. Forensics Security, vol. 15, pp. 578–593, 2020

[11] Z. Cheng and X. Zhang, "Integrating Fine-Grained Classification and Motion Relation Analysis for Face Anti-Spoofing," in IEEE Access, vol. 13, pp. 93649-93660, 2025.

[12] I. Chingovska, A. Anjos, M. Sébastien, "On the effectiveness of local binary patterns in face anti-spoofing," BIOSIG, 2012

[13] Doo-Hyun Choi, I. Jang, Mi Hye Kim, N. Kim, "Color image enhancement using single-scale retinex based on an improved image formation model," 2008, 16th European Signal Processing Conference (2008): 1-5.

[14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, N. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale", 2021 9th International Conference on Learning Representations (ICLR).

[15] Z. Du, J. Li, L. Zuo, L. Zhu, and K. Lu, "Energy-based domain generalization for face anti-spoofing," in Proc. 30th ACM Int. Conf. Multimedia, Oct. 2022, pp. 1749–1757.

[16] A. George, Z. Mostaani, D. Geissenbuhler, O. Nikisins, A. Anjos and S. Marcel, "Biometric Face Presentation Attack Detection With Multi-Channel Convolutional Neural Network," in IEEE Transactions on Information Forensics and Security, vol. 15, pp. 42-55, 2020, doi: 10.1109/TIFS.2019.2916652.

[17] A. George, S. Marcel, "Learning One Class Representations for Face Presentation Attack Detection using Multi-channel Convolutional Neural Networks", in IEEE Transactions on Information Forensics and Security, vol. 16, pp. 361-375, 2021, doi: 10.1109/TIFS.2020.3013214.

[18] A. George and S. Marcel, "Deep pixel-wise binary supervision for face presentation attack detection," in Proc. Int. Conf. Biometrics (ICB), Jun. 2019, pp. 1–8.

[19] A. George and S. Marcel, "On the effectiveness of vision transformers for zero-shot face anti-spoofing," in Proc. IEEE Int. Joint Conf. Biometrics (IJCB), Aug. 2021, pp. 1–8.

[20] H. Gu, J. Chen, F. Xiao, Y. -J. Zhang and Z. -M. Lu, "Self-Attention and MLP Auxiliary Convolution for Face Anti-Spoofing," in IEEE Access, vol. 11, pp. 131152-131167, 2023, doi: 10.1109/ACCESS.2023.3335040.

[21] J. Guo, A. Liu†, Y. Diao†, J. Zhang, H. Ma, B. Zhao, R. Hong, M. Wang, "Domain Generalization for Face Anti-spoofing via Content-aware Composite Prompt Engineering", 2025

[22] G. Heusch, A. George, D. Geissbuhler, Z. Mostaani, S. S. Marcel, "Deep Models and Shortwave Infrared Information to Detect Face Presentation Attacks", in IEEE Trans. Biom. Behav. Identity Sci., vol. 2, pp. 399-409, 2020, doi.org/10.1109/TBIOM.2020.3010312.

[23] M. S. Hossain, L. Rupty, K. Roy, M. Hasan, S. Sengupta, and N. Mohammed, "A-DeepPixBis: Attentional angular margin for face anti-spoofing," in Proc. IEEE Digit. Image Comput. Techn. Appl., 2020, pp. 1–8

[24] X. Huang, J. Xia, S. Linlin, "One-Class Face Anti-spoofing Based on Attention Auto-encoder," in Proc. Biometric Recognition: 15th Chinese Conference, 2021.

[25] P. K. Huang, C. H. Chiang, T. H. Chen, J. X. Chong, T. L. Liu and C. T. Hsu, "One-Class Face Anti-Spoofing via Spoof Cue Map-Guided Feature Learning," IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024.

[26] Q. Hui, R. Han, Y. Shi, Q. Xiaobo, "A Novel High-Performance Face Anti-Spoofing Detection Method", vol. 12, IEEE, pp. 67379–67391, 2024.

[27] J. Kannala, E. Rahtu, "BSIF: Binarized statistical image features," in Proc. International Conference on Image Processing, 2012, pp.1363-1366.

[28] CH. Liao, WC. Chen, HT. Liu, YR. Yeh, MC. Hu, CS. Chen. "Domain invariant vision transformer learning for face anti-spoofing." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023.

[29] S. Lim, Y. Gwak, W. Kim, J. -H. Roh and S. Cho, "One-Class Learning Method Based on Live Correlation Loss for Face Anti-Spoofing," in IEEE Access, vol. 8, pp. 201635-201648, 2020

[30] Y. Liu, J. Stehouwer, A. Jourabloo, and X. Liu, "Deep tree learning for zero-shot face anti-spoofing," in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2019.

[31] A. Liu et al., "Face anti-spoofing via adversarial cross-modality translation," IEEE Trans. Inf. Forensics Security, vol. 16, pp. 2759–2772, 202

[32] A. Liu, S. Xue2, J. Gan3, J. Wan1, Y. Liang4, J. Deng, S Escalera, Z. Lei., "CFPL-FAS: Class Free Prompt Learning for Generalizable Face Anti-spoofing", in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2024.

[33] A. Liu, Z. Tan, Z. Yu, C. Zhao, J. Wan, Y. Liang, Z. Lei, D. Zhang, S. Z. Li, and G. Guo. 2023. "FM-ViT: Flexible Modal Vision Transformers for Face Anti-Spoofing", Trans. Info. For. Sec. 18 (2023), 4775–4786.

[34] A. Liu, H. Yuan, X. Guo, H. Ma, W. Zhuang, C. Miao, Y. Hong, C. Song, J. Lan, Q. Chu, T. Gong, Y. Liang, W. Wang, J. Wan, X. Liu, Z. Lei. "Benchmarking Unified Face Attack Detection via Hierarchical Prompt Tuning", 10.48550/arXiv.2505.13327, 2025.

[35] S. Mehta, M. Rastegari, "MobileViT: Light-weight, General-purpose, and Mobile-friendly Vision Transformer", 2022 International Conference on Learning Representations.

[36] K. Narayan, V. Patel, "Hyp-OC: Hyperbolic One Class Classification for Face Anti-Spoofing", 2024 International Conference on Automatic Face and Gesture Recognition (FG).

[37] P. V. Ngoan and L. H. Trang, "Deep and wide features for face antispoofing," in Proc. 5th Int. Conf. Big Data Internet Things, Aug. 2022, pp. 39–42.

[38] S. M. Nguyen, L. D. Tran, D. V. Le and A. Masayuki, "Self-Attention Generative Distribution Adversarial Network for Few- and Zero-Shot Face Anti-Spoofing," 2022 IEEE International Joint Conference on Biometrics (IJCB), Abu Dhabi, United Arab Emirates, 2022, pp. 1-9, doi: 10.1109/IJCB54206.2022.10007986.

[39] O. Nikisins, A. Mohammadi, A. Anjos and S. Marcel, "On Effectiveness of Anomaly Detection Approaches against Unseen Presentation Attacks in Face Anti-spoofing," 2018 International Conference on Biometrics (ICB), Gold Coast, QLD, Australia, 2018, pp. 75-81, doi: 10.1109/ICB2018.2018.00022.

[40] T. Ojala, M. Pietikäinen, David Harwood, "A comparative study of texture measures with classification based on featured distributions," Pattern Recognit. 29 (1996): 51-59.

[41] P. Perera and V. M. Patel, "Learning deep features for one-class classification," IEEE Transactions on Image Processing, vol. 28, no. 11, pp. 5450–5463, 2019.

[42] D. Pérez-Cabo, D. Jiménez-Cabello, A. Costa-Pazo, and R. J. López-Sastre, "Learning to learn face-pad: a lifelong learning approach," in IJCB. IEEE, 2020.

[43] Y. Qin et al., "Learning meta model for zero-and few-shot face anti-spoofing," in Proc. AAAI Conf. Artif. Intell., 2020, pp. 11916–11923.

[44] P. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis", in Journal of Computational and Applied Mathematics, 1987, pp. 53-65.

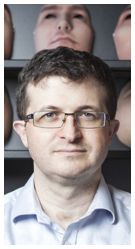
[45] B. Stark, M. Mcgee, Y. Chen, "Short wave infrared (SWIR) imaging systems using small Unmanned Aerial Systems (sUAS)," International Conference on Unmanned Aircraft Systems, ICUAS, 2015, pp.495-501.

[46] W. Wang, F. Wen, H. Zheng, R. Ying, and P. Liu, "Conv-MLP: A convolution and MLP mixed model for multimodal face anti-spoofing," IEEE Trans. Inf. Forensics Security, vol. 17, pp. 2284–2297, 2022.

[47] D. Wen, H. Han and A. K. Jain, "Face Spoof Detection With Image Distortion Analysis," in IEEE Transactions on Information Forensics and Security, vol. 10, no. 4, pp. 746-761, 2015

[48] F. Xiong and W. AbdAlmageed, "Unknown Presentation Attack Detection with Face RGB Images," 2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS), Redondo Beach, CA, USA, 2018, pp. 1-9, doi: 10.1109/BTAS.2018.8698574.
[49] Z. Yu, J. Wan, Y. Qin, X. Li, S. Z. Li and G. Zhao, "NAS-FAS: Static-Dynamic Central Difference Network Search for Face Anti-Spoofing," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 9, pp. 3005-3023.
[50] Z. Yu, Y. Qin, X. Li, C. Zhao, Z. Lei and G. Zhao, "Deep Learning for Face Anti-Spoofing: A Survey," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 5, pp. 5609-5631, 1 May 2023, doi: 10.1109/TPAMI.2022.3215850.
[51] Q. Zhou et al., "Instance-Aware Domain Generalization for Face Anti-Spoofing," 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 2023, pp. 20453-20463.

Anjith George has received his Ph.D. and M-Tech degree from the Department of Electrical Engineering, Indian Institute of Technology (IIT) Kharagpur, India in 2012 and 2018 respectively. After Ph.D, he worked in Samsung Research Institute as a machine learning researcher. Currently, he is a research associate in the biometric security and privacy group at Idiap Research Institute, focusing on developing face recognition and presentation attack detection algorithms. His research interests are real-time signal and image processing, embedded systems, computer vision, machine learning with a special focus on Biometrics.



Sébastien Marcel heads the Biometrics Security and Privacy group at Idiap Research Institute (Switzerland) and conducts research on face recognition, speaker recognition, vein recognition, attack detection (presentation attacks, morphing attacks, deepfakes) and template protection. He received his Ph.D. degree in signal processing from Université de Rennes I in France (2000) at CNET, the research center of France Telecom (now Orange Labs). He is Professor at the University of Lausanne (School of Criminal Justice) and a lecturer at the École Polytechnique Fédérale de Lausanne. He is also the Director of the Swiss Center for Biometrics Research and Testing, which conducts certifications of biometric products.

Igor Bastos has received his Ph.D. degree from the Federal University of Minas Gerais (UFMG), in 2020. After Ph.D, he started to work as Data Scientist in IFood, elaborating machine learning models in areas such as Computer Vision and Natural Language Processing. His research interests are real-time image processing, computer vision and natural language processing, with focus on Activity Recognition and Biometrics.

Anderson Rocha (IEEE Fellow) is Full-Professor of Artificial Intelligence and Digital Forensics at the Institute of Computing, University of Campinas (Unicamp), Brazil. His main interests include Artificial Intelligence, Digital Forensics, Reasoning for Complex Data, and Machine Intelligence. He is the Head of the Artificial Intelligence Lab., Recod.ai, at Unicamp and was the former Director of the Institute for the 2019-2023 term. He is an elected affiliate of the Brazilian Academy of Sciences (ABC) and the Brazilian Academy of Forensic Sciences (ABC). He is a three-term elected member of the IEEE Information Forensics and Security Technical Committee (IFS-TC) and a former chair of such committee. In 2023, he was elected again to the IFS-TC chair for the 2025-2026 term. He has actively been an editor of important international journals and the chair of key conferences in Al and digital forensics. He is a Microsoft Research and a Google Research Faculty Fellow, important academic recognitions given to researchers by Microsoft Research and Google. In addition, in 2016, he was awarded the Tan Chin Tuan (TCT) Fellowship, a recognition promoted by the Tan Chin Tuan Foundation in Singapore. Since 2023, he is also an Asia Pacific Artificial Intelligence Association Fellow. He has been the principal investigator of several research projects in partnership with public funding agencies in Brazil and abroad and national and multinational companies, having already filed and licensed several patents. He is a Brazilian CNPq research scholar (PQ1C). He is ranked among the Top 2% of research scientists worldwide, according to PlosOne/Stanford and Research.com studies. Finally, he is now a LinkedIn Top Voice in Artificial Intelligence for continuously raising awareness of Al and its potential impacts on society