Received: 1 November 2020 | Revised: 16 April 2021 | Accepted: 12 May 2021

DOI: 10.1049/bme2.12045

IET Biometrics

The Institution of Engineering and Technology WILEY

ORIGINAL RESEARCH PAPER

An exploratory study of interpretability for face presentation attack detection

Ana F. Sequeira[1] | Tiago Gonçalves[1,2] | Wilson Silva[1,2] | João Ribeiro Pinto[1,2] | Jaime S. Cardoso[1,2]

[1]INESC TEC Porto, Porto, Portugal

[2]Faculdade de Engenharia da Universidade do Porto, Porto, Portugal

Correspondence

Ana F. Sequeira, INESC TEC, Campus da FEUP, Rua Dr Roberto Frias 4200-465 Porto, Portugal.
Email: ana.f.sequeira@inesctec.pt

Funding information

Fundação para a Ciência e a Tecnologia within project UIDB/50014/2020 and the PhD grants, Grant/Award Number: SFRH/BD/137720/2018; SFRH/BD/139468/2018; and 2020.06434.BD

Abstract

Biometric recognition and presentation attack detection (PAD) methods strongly rely on deep learning algorithms. Though often more accurate, these models operate as complex black boxes. Interpretability tools are now being used to delve deeper into the operation of these methods, which is why this work advocates their integration in the PAD scenario. Building upon previous work, a face PAD model based on convolutional neural networks was implemented and evaluated both through traditional PAD metrics and with interpretability tools. An evaluation on the stability of the explanations obtained from testing models with attacks known and unknown in the learning step is made. To overcome the limitations of direct comparison, a suitable representation of the explanations is constructed to quantify how much two explanations differ from each other. From the point of view of interpretability, the results obtained in intra and inter class comparisons led to the conclusion that the presence of more attacks during training has a positive effect in the generalisation and robustness of the models. This is an exploratory study that confirms the urge to establish new approaches in biometrics that incorporate interpretability tools. Moreover, there is a need for methodologies to assess and compare the quality of explanations.

1 | INTRODUCTION

The success of artificial intelligence (AI) systems demonstrated by the excellence of machine learning (ML)-based systems in most of the AI fields outperforming traditional handcrafted methods, is mainly due to improvements in deep learning (DL) methodology, availability of large databases, and computational gains obtained with powerful graphics processing unit (GPU) cards [1, 2].

One of the greatest current challenges related to AI is the lack of transparency of DL algorithms [3–5]. After the euphoria around artificial neural networks and their over-performing accuracy rates, the research community is starting to understand the drawbacks of black-box algorithms and the importance of being able to understand their decisions and reasoning.

In particular, biometric systems (BS) and anti-spoofing techniques may be positively impacted by the use of interpretability. Most BS can be spoofed by an attacker presenting fake or altered samples of the biometric trait at the sensor. Presentation attack detection (PAD) methods are intended to detect spoofing attacks. When designing a PAD method, it can be very rewarding to know more about the rationale behind its predictions instead of just blindly relying on its outputs. Hence, studying what a model learns and which information it uses to decide about a threat is very beneficial.

As in several other pattern recognition tasks, the use of DL-based PAD methods is increasingly common [6, 7]. There is a pressing need for the use of interpretability as the artificial neural networks become deeper and the process more elusive. Moreover, traditional evaluation setups, whose adequacy has been questioned considering the robustness to unseen attacks [8–10], may not be able to thoroughly capture the model's behaviour. Also, traditional metrics quantify the performance solely relying on predicted labels, without looking into the information used to reach these predictions.

This assessment is quite limited, especially for DL-based approaches.

Considering the aforementioned limitations of the current approach to PAD, this work argues that the evaluation frameworks need to be reformulated to become more thorough and meaningful. Interpretability, with its ability to provide insights into the operation of complex models, can be the key to achieve this goal by providing complementary information. Another perspective is the need to test the robustness of the models and their capacity to generalise to unknown attacks, that is, types of attacks that were not seen by the model during the training phase. Again, through interpretability it may be possible to analyse how differently a model is behaving in the face of a known/unknown (seen/unseen) type of attack sample.

Our previous work [11] offered a preliminary interpretability analysis of a face PAD method. The insights on PAD performance offered by both traditional metrics and a state-of-the-art interpretability tool (Grad-CAM [12]) were subjectively compared to assess whether the latter could offer useful information on the model's behaviour. This work extends the previous research with a more thorough and systematic analysis of the explanations, offering more solid and meaningful conclusions on the potential of interpretability for PAD. The analysis proposed in the current work went beyond the visual inspection of the explanations of a few examples. Inspired by state-of-the-art methodologies, a suitable semantic representation of the explanations was produced such that it was possible to quantify how much the two explanations differ from each other.

It is not a goal of this exploratory study to push forward the state of the art of face PAD methods or address the generalisation problem in itself. The aim is instead to push forward the non-existent field of the explainability analysis of the biometric topics. At the moment, there are no available methodologies to ascertain the performance of a model from the explainable artificial intelligence (xAI) perspective or to assess the quality of the explanations provided. The main contributions of this work are as follows:

1. A pioneering study of interpretability on a face PAD method using data comprising a wide range of attacks (image frames extracted from videos consisting of several types of worn paper masks, paper photos, and replayed recordings);
2. The comparison between traditional performance metrics and interpretability tools in different evaluation frameworks (One-Attack vs. Unseen-Attack);
3. A systematic evaluation of the explanations obtained regarding the PAD models' decisions (for different evaluation frameworks) for both the *bona fide* and the *presentation attack* samples: (a) in intra-class studies, comparing (i) different explanations for the same sample and (ii) explanations for different samples; (b) in inter-class studies, producing an overall comparison.

The remainder of this work is organised as follows: Section 2 describes the concepts of PAD and the face PAD network used; Section 3 presents the related work on interpretability, biometrics and PAD; Section 4 highlights the challenges addressed by the proposed study; Sections 5–8 describe the proposed methodology; Section 9 details the experimental setup; Section 10 presents the results and their discussion; and Section 11 sums up the conclusions drawn from this work and points directions for future work.

2 | PRESENTATION ATTACK DETECTION (PAD)

2.1 | PAD evaluation frameworks

The main definitions regarding presentation attack detection (PAD) used in the work follow the ISO/IEC 30,107-3:2017 standard [13]. The more classical approaches in PAD use only one type of attack, that is, one PAI Species (PAISp) to train and test the model. As in the authors' view, this approach leads to a very optimist evaluation of the classifier's performance. It is necessary to test the robustness of the model to PAISp not shown at the time of training. In the present work, the experiments analysed the explanations obtained in different evaluation frameworks described as follows:

One-Attack: The model is trained and tested with *bona fide* samples and only one type of attack. Therefore, the only type of attack shown to the network during the test phase was already seen in the training step. The expression One-Attack#i means that the respective model was trained and tested with *bona fide* samples and *presentation attack* samples of type i.

Unseen-Attack: The model is trained with all but one type of attack and tested with this remaining attack, besides the *bona fide* samples in the training and testing steps. Therefore, during the test phase the network is evaluated only with one type of attack that was not present in the training step—referred to as the unseen attack. Whereas in the training phase, all the other types of attacks were available. The expression Unseen-Attack#i means that the respective model was tested with *bona fide* samples and *presentation attack* samples of type i and trained with *bona fide* samples and the remaining types of attacks (i.e., trained with $j \in \{1,...,7\} \setminus \ i$).

2.2 | Face presentation attack detection model

A PAD method receives a biometric trait measurement as the input and returns a prediction of a *bona fide presentation* or *presentation attack*. As in [11], the model used for PAD is an end-to-end convolutional neural network (CNN) with a relatively simple architecture, detailed in Figure 1. As an end-to-end CNN, the model could freely learn the most appropriate features for this task, which is the most interesting context on interpretability studies focused on gaining insight into the inner workings of a classifier.

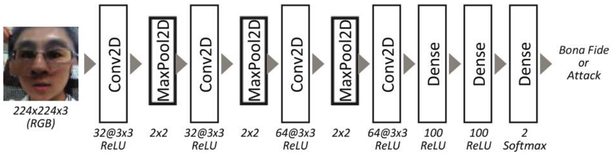


FIGURE 1 Architecture of the implemented presentation attack detection end-to-end deep model (from [11])

3 | RELATED WORK

3.1 | Interpretability concepts and literature

There is no single definition regarding interpretability and explainability in AI. For some authors there is a clear distinction between them, leading to distinct definitions as follows. *Interpretability*: an interpretation is the mapping of an abstract concept (e.g. a predicted class) into a domain that a human can grasp. *Explainability*: an explanation is the collection of features of the interpretable domain that have contributed to the produced decision (e.g. classification or regression) [14]. Other authors use the terms interchangeably, regarding both interpretability and explainability as a three-stage process in the development cycle of an ML model, with these stages being named as pre-, in-, and post-model [5] stages. To date, there is still a predominance of studies putting efforts on this last stage of post-model interpretability, where the focus is on understanding an unconstrained previously built model.

The efforts of the research community posed on the field of xAI resulted in the development of both interpretable models (pre-model and in-model stage) and explanation methods (post-model stage) over the past few years [3, 5, 15–18]. Nonetheless, most of the efforts have been put on these explanation methods where the focus is more on understanding an unconstrained and previously built model than on creating intrinsically interpretable models. The xAI contributions span over the fundamental research in machine learning to applications in other fields such as medicine [16, 19, 20] or finance [17]. It is not a coincidence that the pioneer application fields are ones where it is of huge importance to foster awareness for the advantages and the necessity of transparent decision making.

Now that the dust raised by the euphoria surrounding artificial neural networks and their over-performing accuracy is starting to settle down, the research community is alert to the reality of being made accountable for what these outstanding models actually learn and decide. The consensus is that much can be learnt by understanding the powerful, black-box-like deep learning models that achieve remarkable accuracy but provide no information about what exactly makes them reach their predictions. There is a growing body of work in the literature devoted to interpreting and explaining the behaviour of machine learning systems for various problems [12, 21–23].

3.2 | xAI for biometrics and PAD

PAD methods in general, and face-focused ones in particular, have demonstrated remarkable performances. This is mostly due to the advantages withdrawn from using efficient deep CNN models [24–26]. The PAD generalisation problem has also been addressed in recent works and breakthroughs were accomplished, with one-class classification or anomaly detection approaches [6, 27–32], fostering the robustness of face PAD methods to unknown attacks.

Regarding xAI in the biometrics field, Zee et al. [33] combined face recognition and a face PAD method and tried to use the interpretations to enhance the performance of the recognition method. In the line of thought of going beyond the binary supervision by using only the labels of the two classes, another work used a depth map and the rPPG signal as the auxiliary supervision to improve the performance of the face anti-spoofing method [34]. More recently, the interpretability and explainability of machine learning models have gained relevance and more works are focused on the application of their methodologies in the field of biometrics and, in particular, to the face PAD problem whether by attempting to estimate the depth map [35–37], provide saliency maps in the CNN model [38, 39] or even studies on the estimation of patterns that characterise an attack sample [38, 40]. To the extent of the authors' knowledge, it is still a territory that has been explored very little to apply interpretability tools and analyse biometric recognition and PAD techniques from the xAI perspective. The current work reinforces the authors' idea that the research in PAD methods may be remarkably impacted by the outcomes of following this path. The first effects in the PAD field will be obtained via the reinforcement of trust as an outcome of model validation as well as the improvement of PAD models' robustness through the detection of their hidden vulnerabilities [11]. From the authors' previous work [11], it was proposed that through interpretability it was possible to grasp the limitations of the models in satisfyingly generalising from the training data. These observations highlight the necessity for performing model validation using interpretability tools, taking advantage of the acquired knowledge to interpret the decision-making process and anticipate vulnerabilities and ultimately, to adapt the model to overcome the anticipated vulnerabilities. In [41] are highlighted the current limitations and the need for interpretability in biometrics, noting that for over 25 years researchers in face recognition and biometrics have measured the accuracy of algorithms and systems decision on benchmark datasets (the 'decision accuracy'). However, through the new perspective, regardless of whether the system decision is ultimately correct or incorrect, an explanation must describe how the system came to its conclusion and it should be noted that an accurate explanation does not imply that a system provided the correct answer. The open challenge pointed out by the authors here is that, while the community has established the 'decision accuracy metrics', researchers have not developed performance metrics for explanations. In [42, 43] robust iris PAD methods are analysed from the perspective of interpretability. Focusing on face morphing attacks, Seibold et al. [44]

proposed a new interpretability method for DNN-based morphing attack detectors that determines which regions of an image contain artefacts. Interpretability has also been the focus of study on other topics related to biometrics, including biometric template security [45] and fingerprint segmentation [46].

The need to evaluate the explanations for the decisions made by the model is one of the major open problems in the field of biometrics today. This open challenge requires research effort to provide breakthroughs that will shed light on new paths on the way to better understanding the black box models that have been used rather trustfully. Moreover, this is also motivated by the regulation. For instance, according to the European Union General Data Protection Regulation (EU-GDPR), companies that want to deploy algorithms which use this type of sensitive data to produce high-stake decisions will have to implement 'suitable measures to safeguard the data subject's rights and freedoms and legitimate interests' and be concerned about Article 22, which states that individuals 'have the right not to be subject to a decision based solely on automated processing' [47]. Therefore, for the implementation of these algorithms in daily life, interpretability will play a very important role. Definitely, it is not yet clear what the appropriate methodologies for defining performance metrics for explanations would be. Further efforts should be devoted to reducing subjectivity in the evaluation of explanations through objective metrics and procedures [48]. The present work aims to be a step further on the path of making interpretability tools ready for application in biometrics and PAD.

4 | OVERVIEW OF THE CHALLENGES ADDRESSED WITH THE INTERPRETABILITY ANALYSIS OF FACE PAD

There are several aspects to investigate in the new horizons opened by the xAI perspective over face PAD. In an earlier work by the authors [11], some 'desirable properties' of PAD methods that can be evaluated using interpretability tools were identified: (1) explanations for the same sample should be similar whether or not it is seen during training (data swap); (2) explanations for the same sample should be similar whether or not the model is trained to detect that specific attack (One-Attack vs. Unseen-Attack); (3) explanations should be similar for different samples with the same label (intra-class coherence); (4) explanations should be meaningful (a human would likely use them to provide the same decision). These properties should be verified by a PAD method that is robust, coherent, meaningful, and can adequately generalise to unseen data and attacks. These desirable properties of a robust PAD method and the methodologies to ascertain them unfold a yet unexplored field of research.

The present work tackles some open questions regarding the stability of the explanations produced by the models varying from the evaluation frameworks as well as how much the explanations vary across the two classes, *bona fide* or *presentation attack*. The first aim is related to the above-mentioned property, 'explanations for the same sample should be similar whether or not the model is trained to detect that specific attack (One-Attack vs. Unseen-Attack)'. It is to be investigated how much the presence or absence of the types of attacks in the training data will affect the way the model learns. Another aim is related to the property, 'explanations should be similar for different samples with the same label (intra-class coherence)'. It is a goal to investigate the intra-class coherence of explanations for the *bona fide* samples and how much the models are affected by variations in the *presentation attack* known in the learning phase. Ideally, a robust PAD method would learn the discriminative features of the *bona fide* samples no matter what attack samples are shown to it. A question is then posed to understand how much the information to classify the *bona fide* samples is affected by changing it between the One-Attack#*i* and One-Attack#*j* frameworks and then between the Unseen-Attack frameworks. To try and answer this, a thorough investigation is performed regarding the *bona fide* samples.

The example shown in Figure 2, analysed in the previous work [11], gave valuable insights into the question posed. It can be observed that although the *bona fide* samples are present in the training step for all these models, the information that the classifier is using to correctly label them varies significantly when the types of *presentation attack* samples in the training step are different. This is more evident for Attack#3: in the case of One-Attack#3 the model was trained and tested with this attack and in the Unseen-Attack#3 it was trained with all but this attack and as a result, the explanations for the classification of the *bona fide* sample differ enormously.

Despite the analysis based on the examples, it is of utmost importance to perform a more systematic study of these matters. Thus, the approaches presented in this work will be applied to perform more than a simple qualitative and somehow subjective analysis. Instead, it is intended to analyse all the *bona fide* images and make a quantitative evaluation of these types of discrepancies between the explanations obtained in the different evaluation frameworks.

To study how the explanations for the *bona fide* images are affected by the presence of different types of attacks in the training set, the set of *bona fide* images correctly classified across all the experiments were selected and the explanations were obtained (in the test step) for both frameworks: One-Attack and Unseen-Attack. The same type of reasoning can be done regarding the *presentation attack* samples. In Figure 3, previously analysed in [11], it can be observed that the information that the classifier is using to correctly label the samples varies significantly depending on whether this type of *presentation attack* samples is present in the training step or not.

The challenges of this work are increased by the lack of suitable metrics and the subjectivity in the evaluation of the explanations. The proposed study aims at being one step further in applying interpretability tools to biometrics and PAD. It should be stressed that improving the state of the art of face PAD methods or the PAD generalisation problem is not the focus of this work. Instead, the aim is to fill the void in the

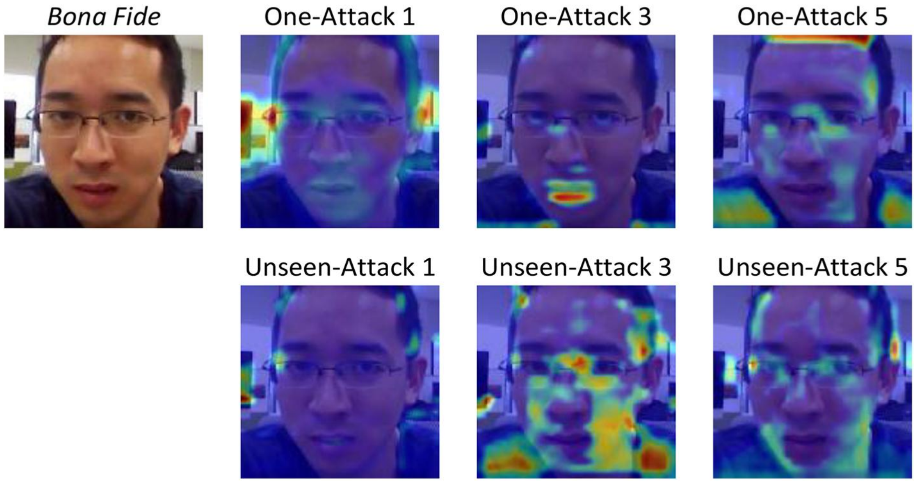


FIGURE 2 Explanations for correctly classified *bona fide* samples (TN) for One-Attack and Unseen-Attack: #1, 3, 5 (from [11])

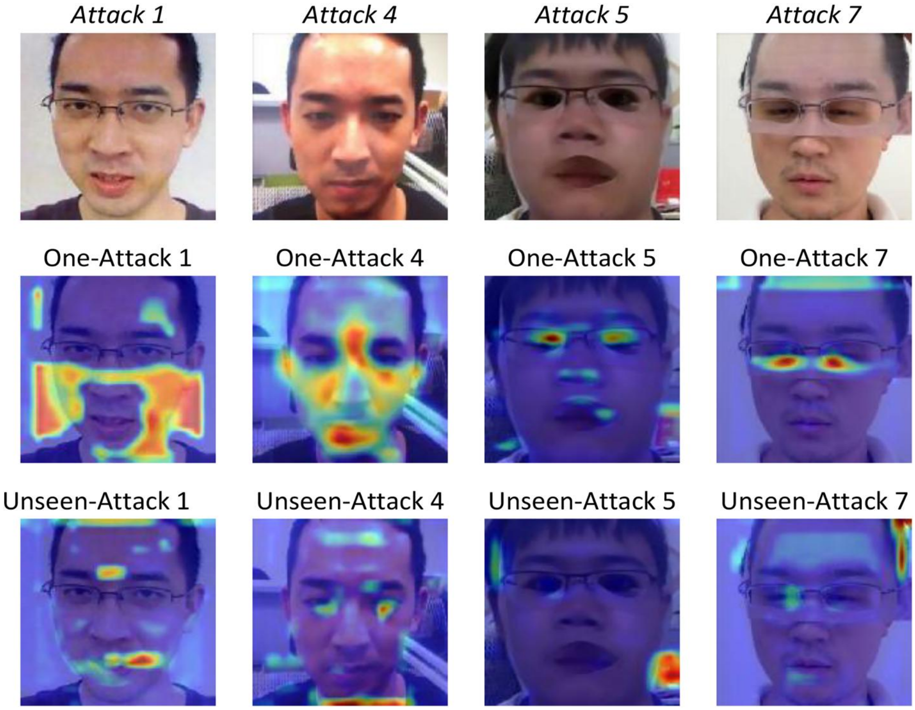


FIGURE 3 Explanations for correctly classified *presentation attack* samples (true positives) for One-Attack and Unseen-Attack: #1, #4, #5 and #7 (from [11])

methodologies to ascertain the performance of a biometrics model and its vulnerabilites from the xAI perspective and ultimately, to assess the quality of the explanations provided.

5 | METHODOLOGY FOR THE REPRESENTATION OF THE PAD MODELS' EXPLANATIONS

5.1 | PAD models' explanations

The quantitative analysis of this work consists in comparing the explanations obtained for the two types of samples (*bona fide* or *presentation attack*) and taking into account the different evaluation frameworks (One-Attack or Unseen-Attack). *Bona fide* images are always present in the training and testing of any framework. Thus, the *bona fide* samples can be tested and an explanation is produced in any framework, One-Attack#i or Unseen-Attack#i.

The *presentation attack* samples belong to a specific type of attack. Considering the One-Attack framework's evaluation, an attack sample of type #i can only be tested for the respective OneAttack#i. It is not meaningful to test this sample with the models of Attack#j (with j $6 = i$) because then, this would be an Unseen-Attack#i scenario.

Let $I = \{I_1,...,I_n\}$ and $E^{xi} = \{E^{xi}_1,\cdots,E^{xi}_n\}$ be a set of images and the respective set of explanations. For each image I_k (for $k = 1,...,n$) there is a corresponding explanation E^{xi}_k obtained, as will be described in section 5.2. Note that $x = o$ or $x = u$ whether the explanation refers to a classification result within the framework One-Attack or Unseen-Attack, respectively, and $i = 1,...,7$ is the type of attack that defines the model used for testing. Thus, each explanation is obtained with a specific model determined by the evaluation framework and the attack used in testing.

5.2 | Semantic representation of explanations

As previously pointed out, this work aims to measure the variability of the explanations for both *bona fide* and *presentation attack* samples in the two different evaluation frameworks. To do this, one needs to define a suitable representation of the explanations produced such that it is possible to quantify how much the two explanations differ from each other. The comparison made and the differences measured are in a semantic context. An illustrative example of how the approach used to perform a quantitative comparison between the explanations is depicted in Figure 4.

The type of explanations considered here are the same as those considered in the authors' earlier work [11]. Thus, the explanations are generated by the Grad-CAM interpretability method [12], which highlights the regions of the image that maximise the predicted class. Since Grad-CAM produces blobby and coarse explanations that highlight the regions without preserving details, in this work it was decided to multiply the saliency maps by the image. Nonetheless, this space is still not ideal for image comparison as this comparison would be highly impacted by the spatial location of important features. To overcome this issue, and inspired by what is being done in image retrieval [49, 50] and concept-based interpretability [51] to find similar images, the learnt features computed by a pre-trained CNN were used as the space to measure the distance between two explanations. This follows the finding of Zhang et al. [52] that the Euclidean distance in the activation space of final layers is an effective similarity metric. Since this work uses face images instead of using a typical ImageNet based pre-training of a CNN, which is adapted to natural images, a face-specific network, FaceNet [53], pre-trained in the VGGFace2 dataset [54], was used for the extraction of deep features. This deep convolutional neural network was trained using a triplet loss. The main goal was to optimise the embedding space and ensure that FaceNet could learn a function that correctly maps the face images to a compact Euclidean space where distances directly correspond to a measure of facial similarity. To take advantage of these FaceNet properties and to achieve meaningful mappings of the Grad-CAM explanations, we start by multiplying the original image by its Grad-CAM explanation. We then input this resulting image into FaceNet[1] (pre-trained on the VGGFace2 dataset). We then extract the features generated in the second to last layer of FaceNet. All the Euclidean distances reported in this work are computed in this semantic space.

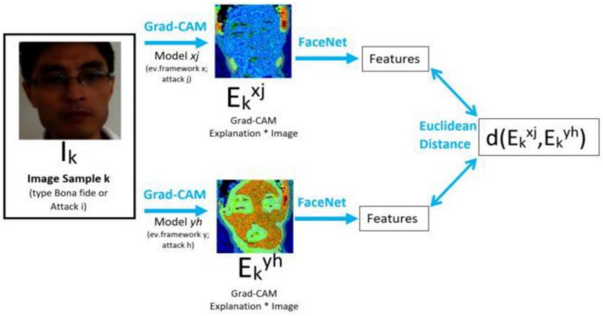


FIGURE 4 An illustrative example of the approach used to quantify how much the two explanations differ from each other

6 | METHODOLOGY FOR THE QUANTITATIVE COMPARISON OF PAD EXPLANATIONS FOR THE SAME SAMPLE IN DIFFERENT EVALUATION FRAMEWORKS

This section describes the study that addresses the property which states that for a robust PAD model, 'explanations for the same sample should be similar whether or not the model is trained to detect that specific attack (One-Attack vs. Unseen-Attack).

Figure 5 illustrates the process of comparing the explanations with respect to one *bona fide* image, I_k and the evaluation framework x (either One or Unseen-Attack) and fixing as reference the explanation obtained by the model for the Attack#i in both frameworks.

Figure 6 illustrates the process of comparing the explanations regarding one *presentation attack* image I_k of type Attack#i. In this case, the comparison is made by fixing as reference the explanation of the result of the classification in the One-Attack framework obtained by the model for the Attack#i; thus, the evaluation framework is $x = o$. As mentioned before, this is done in order to have a more stable benchmark for comparison, since in the One-Attack framework the model is trained and tested with the same type of attack.

So, for each image I_k (either *bona fide* or *attack*), using the evaluation framework x and the model regarding *Attack#i*, a set of 6 values $\{d^{xj}{}_k: j \in \{1,...,7\} \setminus i\}$ is obtained by comparing the explanation E^{xi}_k (always E^{oi}_k in the PA case) with the explanation $E_k{}^{xj}$ (for $j \in \{1,...,7\} \setminus i$). These distance measurements between explanations $\{d^{xj}{}_k\}$, obtained as described in section 5.2, provide a quantitative measure of the variability of the explanations produced by the different models (trained in

[1] Available at: https://github.com/timesler/facenetpytorch

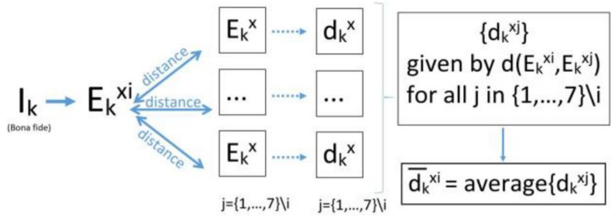


FIGURE 5 Comparison of explanations for *bona fide* samples, image I_k, evaluation framework x and fixing Attack#i

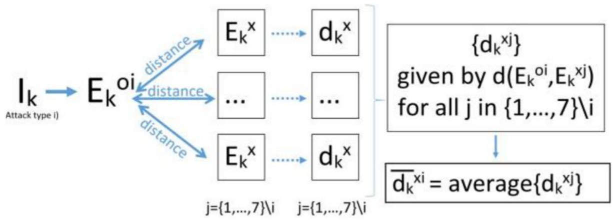


FIGURE 6 Comparison of explanations for the *presentation attack* sample, image I_k, evaluation framework x and fixing Attack#i

different conditions using different attack types) regarding the same image. Averaging these values will provide the $\bar{d}_k^{xi}$ values for comparison.

Let the average of the $\{d^{xj}{}_k\}$ values, $\bar{d}_k^{xi}$, be given by Equation (1):

$$\bar{d}_k^{xi} = \frac{1}{6}\sum_j d_k^{xj} \tag{1}$$

For the sake of clarity, Table 1 presents these distance values and their correspondence to the images and frameworks.

In the following sections, other measures will be obtained from the values $\bar{d}^{xi}{}_k$, obtained for each image I_k. These values are unique for an *attack* sample and multiple (one for each i ∈{1,...,7}) for a *bona fide* sample.

6.1 | Image Average: bona fide and presentation attack

The *Image Average (Iμ)* provides a quantitative measure of the variability (across all models) of the explanations produced by each model under the evaluation framework defined by xi (for $x = o$ or $x = u$ and i = 1,...,7) regarding the image I_k. The $I\mu$ for the *bona fide* is given by Equation (2) and the $I\mu$ for the attacks is given by Equation (3).

$$Bona\,fide : I\mu_k^x = \frac{1}{7}\sum_{i=1}^{7}\bar{d}_k^{xi} \tag{2}$$

$$Attack\,(type\ \#i) : I\mu_k^x = \bar{d}_k^{xi} \tag{3}$$

TABLE 1 Average values obtained from the comparison of explanations for different classes and evaluations frameworks (BF stands for *bona fide*; PA for *presentation attack*)

Class	Framework	Comparisons
BF	One-Aattack	$\{\bar{d}_k^{xi} : \bar{d}_k^{xi} for\ i = 1, \ldots, 7\}$
	Unseen-Attack	$\{\bar{d}_k^{xi} : \bar{d}_k^{xi} for\ i = 1, \ldots, 7\}$
PA (type #i)	One-Attack	$\bar{d}_k^{xi}$
	Unseen-Attack	$\bar{d}_k^{xi}$

6.2 | Attack Average: bona fide and presentation attack

The *Attack Average (Aμ)* provides a quantitative measure of the variability (across all samples) of the explanations produced by the model under the evaluation framework defined by xi (for $x = o$ or $x = u$ and i = 1,...,7). Consider the values $\bar{d}_k^{xi}$ as defined in Table 1 and n and m as the number of *bona fide* and *attack* samples, respectively. The $A\mu$ for the *bona fide* is given by Equation (4) and the $A\mu$ for the attacks is given by Equation (5).

$$Bona\,fide : Au^{xi} = \frac{1}{n}\sum_{i=1}^{n}\bar{d}_k^{xi}, for\ i = 1, \ldots, 7 \tag{4}$$

$$Attack\,(I_k type\ \#i) : A\mu^{xi} = \frac{1}{m}\sum_{i=1}^{m}\bar{d}_k^{xi}, for\ i = 1, \ldots, 7 \tag{5}$$

7 | METHODOLOGY FOR INTER-CLASS COMPARISON IN UNSEEN-ATTACK EVALUATION: BONA FIDE VS PRESENTATION ATTACK

This section describes the study to investigate the inter-class comparison of explanations obtained using the models in the Unseen-Attack framework. In other words, the variability of the explanations between the *bona fide* and *presentation attack* samples is investigated.

To achieve the desired goal, the explanations obtained from the classification of each image will be compared in a pairwise manner with the explanations of the different models trained in the Unseen-Attack framework. This comparison is performed for all images within each class.

By using the Unseen-Attack models it is possible to test the robustness of the models to the variability in the attacks present in the training and testing steps. Recall that a model resulting from Unseen-Attack#i is trained with attacks j for j∈{1,...,7} \ i.

Figure 7 illustrates the process of obtaining the comparison of all explanations of one *bona fide* image I_k. It shows one example of how to obtain the pairwise distances D_k given by D_k = $\{d^{jh}{}_k : d^{jh}{}_k = d(E_k{}^{uj}, E_k{}^{uh})$, with j,h ∈{1,...,7} and j $6 = h\}$. Then the values in D_k are averaged and $\bar{d}_k$ is obtained for image I_k. A global value is obtained averaging all these values, $\bar{d}_{BF}$, as given by Equation (6):

$$\bar{d}_{BF} = \frac{1}{n} \sum_k \bar{d}_k \qquad (6)$$

for the *bona fide* images I_k (k = 1,...,n).

Figure 8 shows the process of obtaining the comparison between the explanations of one *presentation attack* image I_k (of the type Attack#i) obtained for the Unseen-Attack#j models for $j \in \{1,...,7\} \setminus i$. It shows one example, regarding image I_k of type *Attack#i*, of how to obtain the pairwise distances $D_k = \{d_k^{jh} : d_k^{jh} = d(E_k^{uj}, E_k^{uh}), with\, j, h \in \{1, \ldots, 7\} \setminus i\} and\, j6 = h\}$. The values in D_k are averaged and $\bar{d}_k$ is obtained for image I_k. A global value is obtained averaging all these values, $\bar{d}_{PA}$, as given by Equation (7):

$$\bar{d}_{PA} = \frac{1}{m} \sum_k \bar{d}_k \qquad (7)$$

for the *presentation attack* images I_k (k = 1,...,m) of type #i.

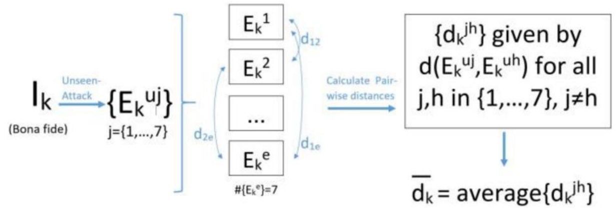
8 | METHODOLOGY FOR COMPARISON OF PAD EXPLANATIONS FOR SAME-CLASS DIFFERENT SAMPLES

This section addresses the property which claims that for a robust PAD model 'explanations should be similar for different samples with the same label'. So, on the one hand it is a goal to investigate the coherence of explanations for the *bona fide* samples. On the other hand, it is a goal to understand how much the explanations vary for *presentation attack samples*. This analysis can be done by observing how much the explanations for the models' decisions are affected by variations in the types of *presentation attacks* known in the learning phase. The comparison is done by comparing features extracted from the explanations and not in a pixel to pixel manner, as described in section 5.2, making it possible to compare the explanations obtained for different samples.

9 | EXPERIMENTAL SETUP

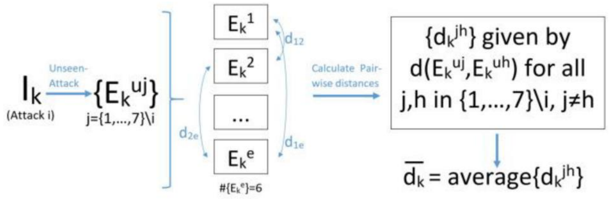
9.1 | Data and pre-processing

The data used was drawn from the ROSE-Youtu Face Liveness Detection Dataset [55]. This dataset is composed of 3497 videos from 20 subjects, including 'genuine' and 'attack' videos. Table 2 details the attack types and the total number of frames for each attack. From each video, frames were extracted every 5s and faces were detected using an MTCNN [56]. Face regions were cropped, resized to 224 × 224, and normalised to [0,1]. The samples from subjects {2,3,4,5,6} were reserved for testing, while the data from the remaining 15 subjects were used for training and validation.



I_k
(Bona fide)
Unseen-Attack
$\{E_k^{uj}\}$
j={1,...,7}
E_k^1
E_k^2
...
E_k^e
#{E_k^e}=7
d_{12}
d_{2e}
d_{1e}
Calculate Pair-wise distances
{d_k^{jh}} given by d(E_k^{uj},E_k^{uh}) for all j,h in {1,...,7}, j≠h
$\bar{d}_k$ = average{d_k^{jh}}

FIGURE 7 Pairwise comparison of explanations produced by the Unseen-Attack models for a *bona fide* sample, I_k



I_k
(Attack i)
Unseen-Attack
$\{E_k^{uj}\}$
j={1,...,7}\i
E_k^1
E_k^2
...
E_k^e
#{E_k^e}=6
d_{12}
d_{2e}
d_{1e}
Calculate Pair-wise distances
{d_k^{jh}} given by d(E_k^{uj},E_k^{uh}) for all j,h in {1,...,7}\i, j≠h
$\bar{d}_k$ = average{d_k^{jh}}

FIGURE 8 Pairwise comparison of explanations produced by the Unseen-Attack models for the *attack* sample, I_k, of type #i

9.2 | Evaluation metrics

The metrics used for the evaluation of PAD models are as follows: the *Bona fide Presentation Classification Error Rate (BPCER)* (the proportion of *bona fide* presentations erroneously classified as attacks); and the Attack Presentation Classification Error Rate (APCER) (the proportion of presentation attack wrongly classified as bona fide) [13]. The Equal Error Rate (EER) is the error at the operation point where the APCER and BPCER take the same value.

9.3 | Implementation details

Using the data previously described, the PAD end-to-end model was trained using the Adam optimiser with an initial learning rate of 0.0001 for a maximum of 150 epochs and batch size 8. Early stopping, dropout, and data augmentation were used as detailed in [11]. All the experiments were performed using the Grad-CAM implementation of the *Keras Visualization Toolkit* [57]. In the saliency maps, each pixel takes a colour from blue to yellow, corresponding to the increasing activation/importance of the pixel.

10 | RESULTS AND DISCUSSION

10.1 | Performance of the face PAD method

Although the focus of this work is on interpreting the decisions of the face PAD model rather than its performance, one should not attempt to interpret a model that lacks PAD capabilities in the first place. Table 3 presents the performance results for One-Attack and Unseen-Attack frameworks.

TABLE 2 Types of presentation attack instruments in the ROSE Youtu DB (N.I. stands for 'number of images', i.e. frames extracted from the videos)

Attack	Types of presentation Attack Instruments	N.I.
-	Genuine (*bonafide*)	2794
#1	Still printed paper	1136
#2	Quivering printed paper	1188
#3	Video of a Lenovo LCD display	923
#4	Video of a Mac LCD display	1113
#5	Paper mask with two eyes and mouth cropped out	608
#6	Paper mask without cropping	1194
#7	Paper mask with the upper part cut in the middle	1162

TABLE 3 PAD performance of the models for One-Attack and Unseen-Attack evaluation frameworks (EER, APCER, and BPCER in %; APCER and BPCER calculated for a threshold of 0.5) (from [11])

	One-Attack			Unseen-Attack		
Attack	*EER*	*APCER*	*BPCER*	*EER*	*APCER*	*BPCER*
1	7.29	12.15	3.06	5.90	6.94	4.90
2	3.62	6.67	1.35	5.55	3.00	10.65
3	2.79	8.37	0.12	10.38	26.29	4.28
4	12.66	30.38	1.84	25.34	45.73	3.92
5	1.61	1.61	1.59	4.84	3.55	7.10
6	4.46	5.10	1.10	10.19	12.74	7.71
7	0.73	5.23	0.00	15.49	34.31	7.71

Notes: Attack Presentation Classification Error Rate (APCER) (the proportion of presentation attack wrongly classified as bona fide) [13]. The Equal Error Rate (EER) is the error at the operation point where the APCER and BPCER take the same value.

Abbreviations: APCER, Attack Presentation Classification Error Rate; BPCER, Bona fide Presentation Classification Error Rate; EER, Equal Error Rate; PAD, presentation attack detection.

10.2 | Comparison of explanations for the same sample in different evaluation frameworks: Image Average (Iμ)

The results presented and discussed in this section come from the quantitative analysis detailed in Section 6. It is an objective of this work to investigate the behaviour of the models when the diversity of attacks in the training data varies. In an ideal scenario of a robust PAD method, the explanations for the same *bona fide* sample (with *bona fide* predicted label) should be similar whether or not the model is trained to detect a specific attack. However, in reality the models are often sensitive to variations in the training data. The same is true for the *presentation attack* samples, where the presence or absence of some attacks affects the behaviour of the models.

Figure 9 shows the mean and standard deviation of the values of *Image Average (Iμ)* values for the experiments described in subsection 6.1. The top image represents the mean value of the *Image Average* values *(Av(Iμ))* and the bottom image represents the standard deviation associated with the mean value of the *Image Average (StD(Iμ))*, for the two types of samples and in the two evaluation frameworks: One-Attack and Unseen-Attack.

From the analysis of Figure 9 several conclusions can be drawn.

First, the intra-class variability is higher in the context of One-Attack than in the context of Unseen-Attack, as indicated by a higher *Av*(*Iμ*) value, regardless of whether *bona fide* or *presentation attack* samples are involved. Moreover, the *Iμ* values within One-Attack show higher variability than within Unseen-Attack for *bona fide*, which can be inferred by the higher *StD*(*Iμ*) value of the One-Attack framework. This suggests that in the Unseen-Attack, the model can better generalise the *bona fide* samples' knowledge because it sees more attack variety during training. For the attack, the intra-class comparison (*bona fide* and *presentation attack*) in both frameworks One-Attack and Unseen-Attack. For the attack samples, the variability among the different types of One-Attack is also higher than the variability of the different types of One-Attack against the different types of Unseen-Attacks. This suggests that even for the *presentation attack* samples, the models become more robust when seeing a bigger diversity of attacks during training.

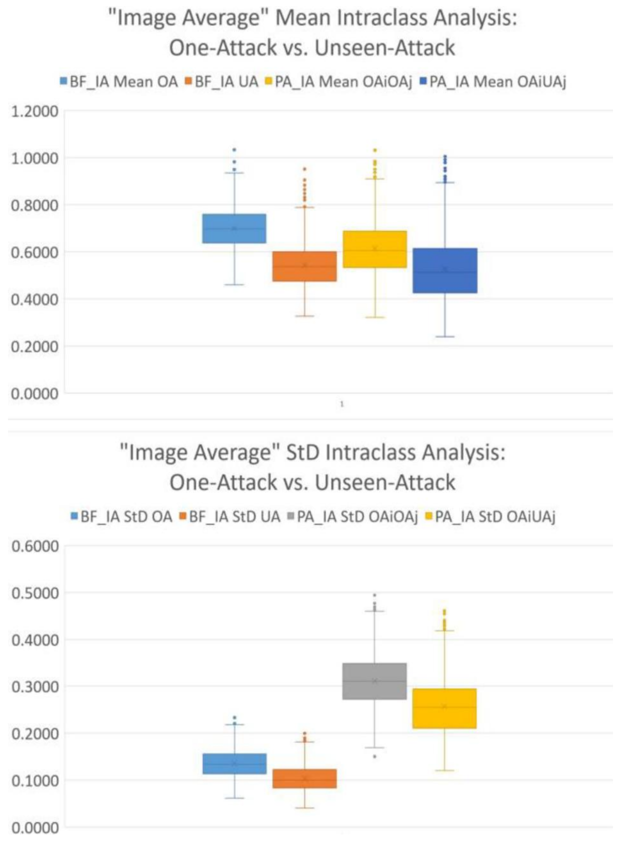


FIGURE 9 Image average mean and standard deviation (StD) for the intraclass comparison (bona fide and presentation attack)in both frameworks One-attack and Unseen-Attack

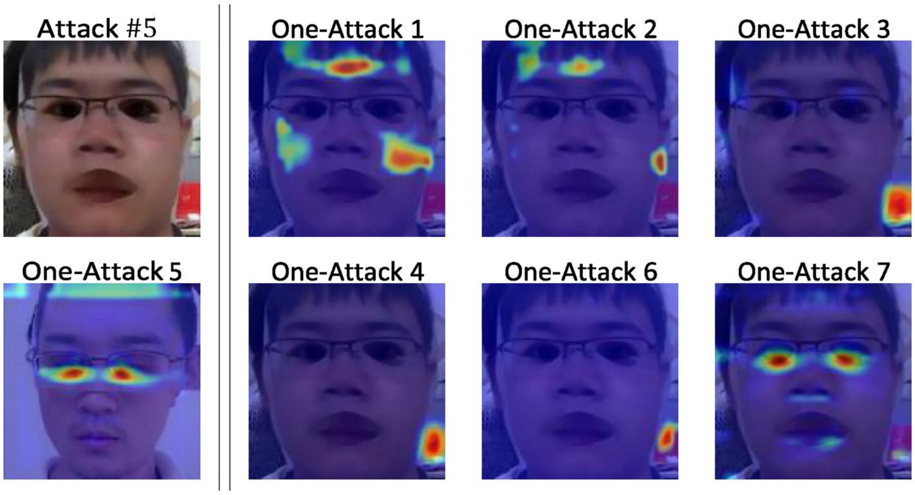


FIGURE 10 Comparison of explanations in intra-class One-Attack: a PA sample of type provides a high Image Average value (obtained when the One-Attack#5 is compared against the One-Attacks {#1,...,#7}\ #5)

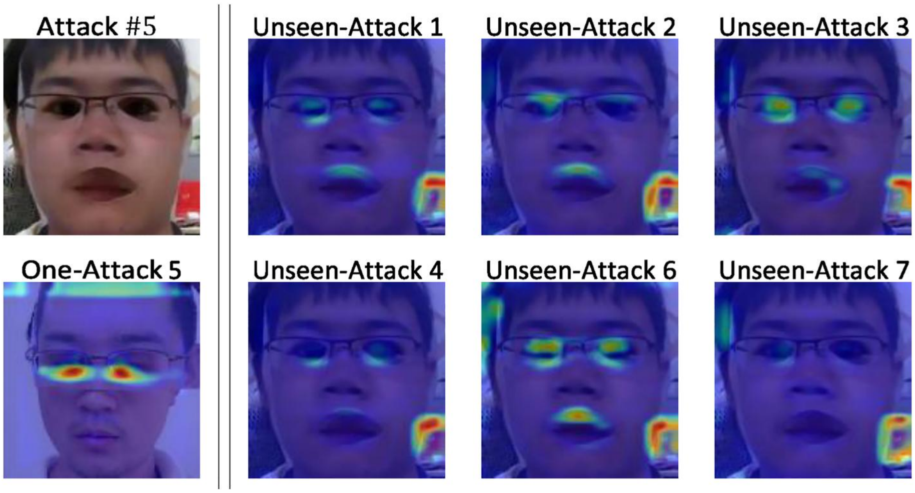


FIGURE 11 Comparison of explanations in intra-class Unseen-Attack: a PA sample of type provides a low *Image Average* value (obtained when the One-Attack#5 is compared against the Unseen-Attacks {#1,...,#7}\ #5)

Figure 10 and Figure 11 are showing an example of an attack sample image that has a comparatively higher $Av(I\mu)$ value in the One-Attack framework and a lower value regarding the Unseen-Attack framework. These results reinforce our intuition that a model that is trained with more than one example of a *presentation attack* ends up learning common patterns that may contribute to a more robust model, with better generalisation ability.

10.3 | Comparison of explanations for the same sample in different evaluation frameworks: Attack Average (Aμ)

The results presented and discussed in this section were obtained from the quantitative analysis detailed in section 6.

Figure 12 and Figure 13 show the Mean of the *Attack Average values* ($Av(A\mu)$), as described in section 6.2, for the *bona fide* and *presentation attack* samples, respectively, for both One-Attack and Unseen-Attack frameworks.

A comparison on the basis of *Attack Average* shows a higher variability for One-Attack, suggesting once again that a training setting that integrates more than one attack may promote the learning of more coherent features for the *bona fide* class (Figure 12). The same is true for the *presentation attack* samples (Figure 13), as the mean distance for the Unseen-Attack framework is smaller than the one for the One-Attack framework. Even though the attacks contain intrinsic specificity, the models used to detect them seem to benefit from the integration of more attacks during the training phase. However, it is interesting to note the variability observed for the *presentation attack* samples' case across the different attacks in the Unseen-Attack framework:

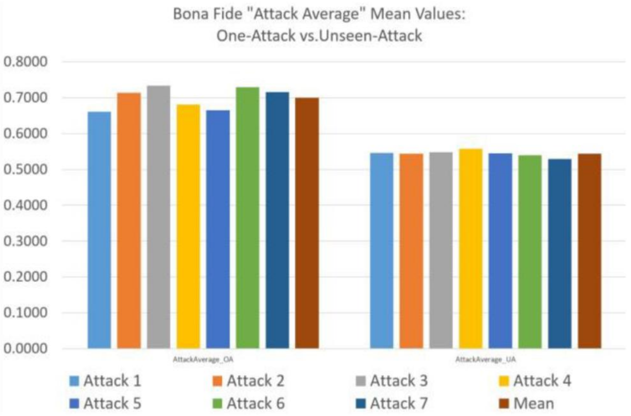


FIGURE 12 Bona fide Attack Average mean for One-Attack and Unseen-Attack (i = 1,...,7) and respective mean values

despite the fact that, in general, the values are smaller than those for the One-Attack (and therefore the models are learning more and becoming more robust to unknown attacks), some specific attacks have much lower values than others. This may mean that the differences in the types of attacks seen in training originate models that are much more sensitive to the unseen data in the testing step than others. In particular, in Figure 13: (i) the Unseen-Attack#5 (consisting of paper masks with eyes and mouth cut out) has a lower value, which can be explained by the fact that in this framework, the model has seen in training many varieties of print attacks (like complete photos, complete paper masks, and half-paper masks); therefore it will generalise more easily; (ii) at the other extreme, the Unseen-Attack#7 (consisting of upper half-paper masks) has a higher value, which can be explained by the fact that these samples combine skin and paper in the facial area; therefore they are more difficult for the model to generalise from the training samples.

10.4 | Inter-class comparison: bona fide versus presentation attack analysis

The results presented and discussed in this section were obtained from the quantitative analysis detailed in section 7.

For the *bona fide* samples, the value $\bar{d}_{BF} = 0.54$ is obtained by averaging the value, $\bar{d}_k$, of all *bona fide* images, I_k (with k = 1,...,n), which is obtained from a pairwise comparison of the explanations of all Unseen-Attack models. The associated standard deviation is 0.13.

Regarding the *presentation attack* samples, the value $\bar{d}_{PA}$ = 0.52 is obtained by averaging the value, $\bar{d}_k$, of all attack images, I_k, (with k = 1,...,m) which is obtained from a pairwise comparison of the explanations of Unseen-Attack#j for $j \in \{1,...,7\} \setminus i$ models. The associated standard deviation is 0.14.

These values are very close and the standard deviation values are not only high but also similar in both cases. The similarity of the values does not allow us to draw a comparative conclusion. Nevertheless, it is worth investigating and trying to interpret the meaning of these absolute values.

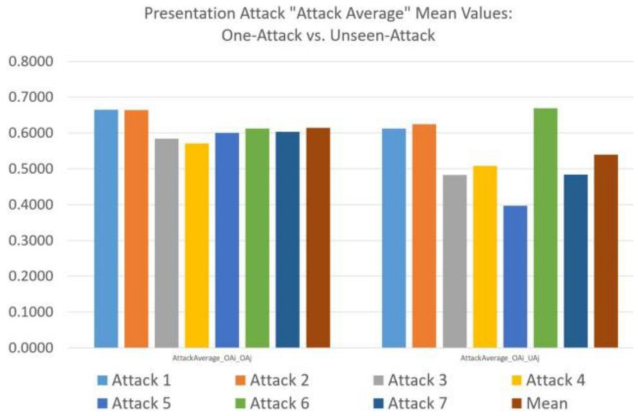


FIGURE 13 Presentation Attack Average mean for One-Attack and Unseen-Attack (i = 1,...,7) and respective mean values

Figure 14 and Figure 15 depict examples of images whose pairwise distance values, $\bar{d}_k$, are close to the calculated mean values, $\bar{d}_{BF}$ and $\bar{d}_{PA}$. By visually inspecting these sets of explanations, it is observed that there is in fact a wide variability between explanations (e.g., in Figure 14 Unseen-Attacks#4,#6 in comparison to #2,#3,#5; and in Figure 15 Unseen-Attacks#3,#4 in comparison to #1,#6).

However, despite the observed variability, in both types of samples there are always certain zones of images that are used by the models to make their decisions. This could be evidence for the fact that, despite the variety of training conditions and the resulting noise, the model is always able to pinpoint some regions of the face that correspond to the real underlying information on the *bona fide* and *presentation attack* labels. This idea is verified when the pairwise distance is above average (Figure 16 and Figure 17).

Obviously, these rationales are based on subjective visual evaluations, but this is a result of the current unavailability of a ground truth for a good or meaningful explanation. These are still muddy grounds that require further research and a combination of interpretability methods with human expert knowledge.

10.5 | Intra-class analysis

The results presented and discussed in this section were obtained from the quantitative analysis detailed in section 8.

In Figure 18, we observe that the One-Attack framework—both for *bona fide* and *presentation attack* samples—leads to higher overall variability. This confirms that a model with a great variety of attacks in the training step will be more consistent in the regions used for the model's decisions.

11 | CONCLUSIONS AND FUTURE WORK

In this work, an analysis of the explanations produced for different models in face PAD was performed. Both

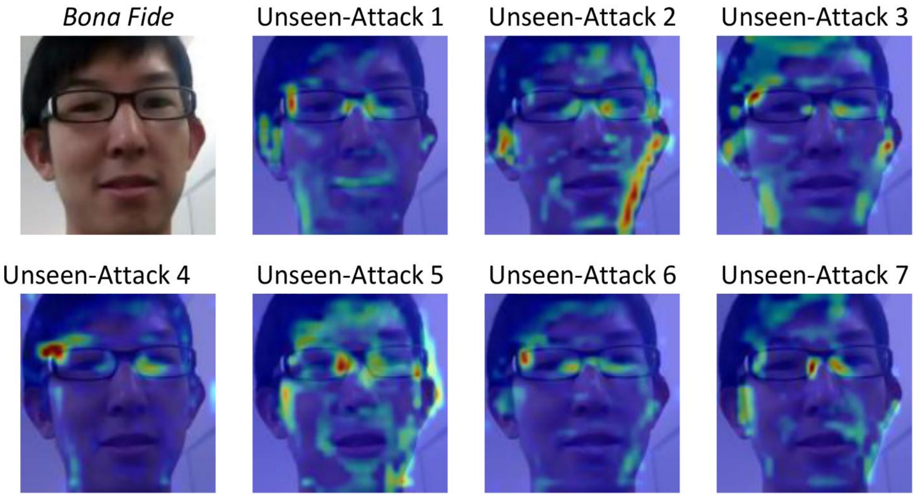


FIGURE 14 Explanations for *bona fide* samples with pairwise distance close to the obtained average ($\bar{d}_{BF} = 0.54$)

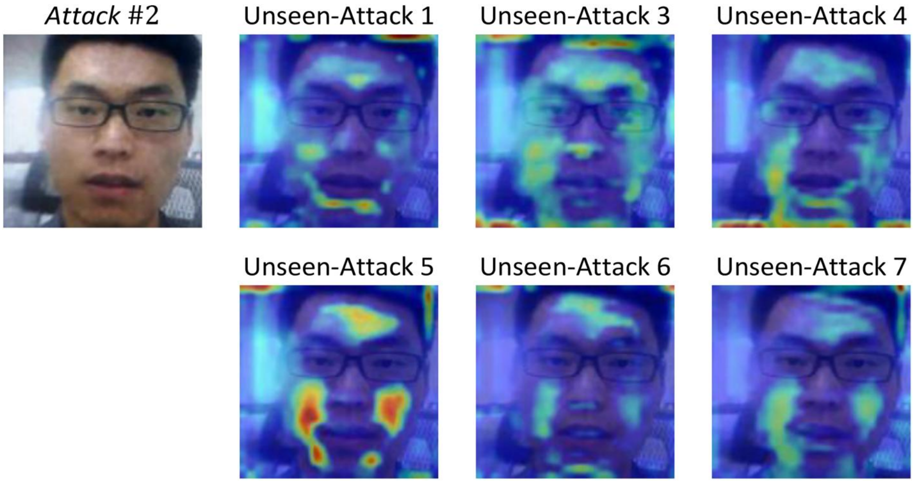


FIGURE 15 Explanations for *presentation attack* sample of type #2 with with pairwise distance close to the obtained average ($\bar{d}_{PA} = 0.52$)

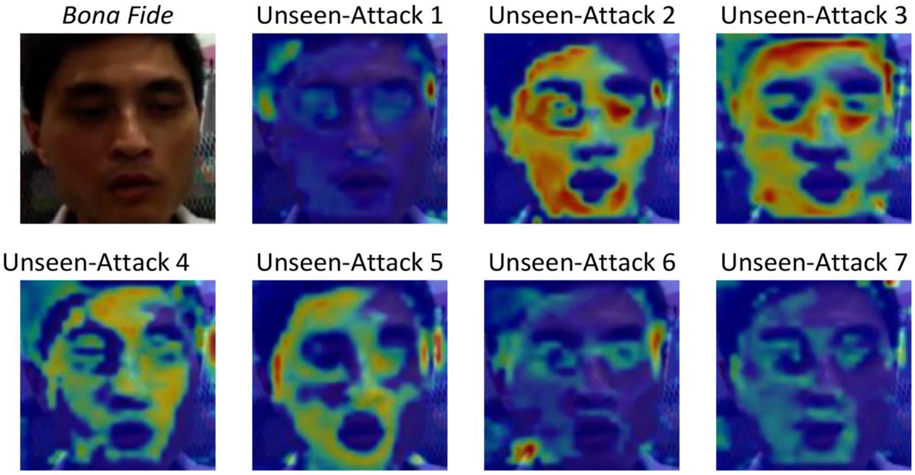


FIGURE 16 Explanations for *bona fide* samples with pairwise distance above the obtained average ($\bar{d}_{BF} = 0.54$)

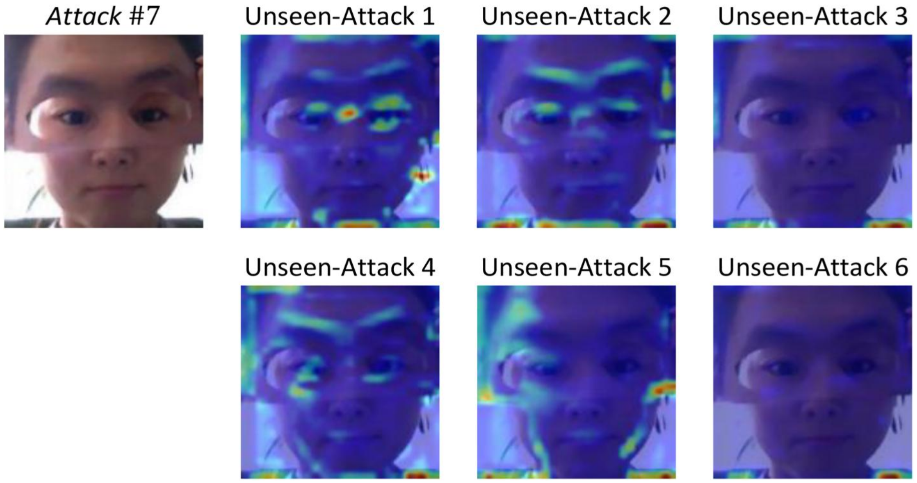
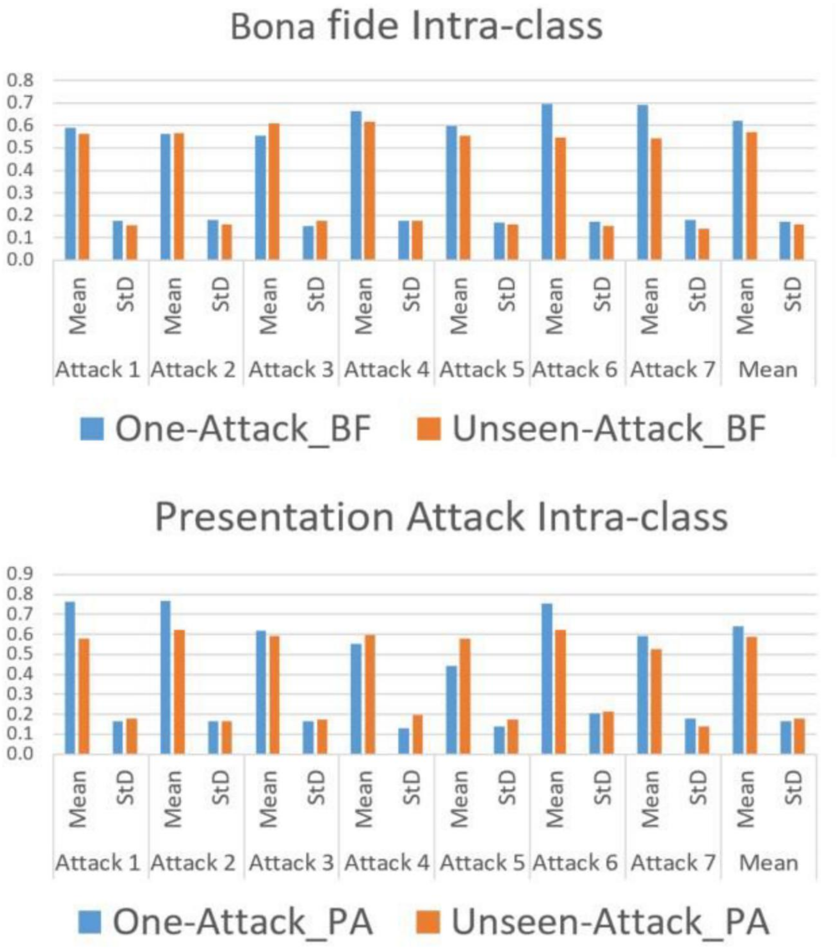


FIGURE 17 Explanations for presentation attack sample of type #7 with pairwise distance above the obtained average ($\bar{d}_{PA} = 0.52$)

FIGURE 18 Bona fide and Presentation Attack Intra-class comparison of mean and std values of One-Attack and Unseen-Attack ($i = 1,...,7$) and respective overall mean/std values



One-Attack and Unseen-Attack training frameworks were considered to analyse the explanations' variability for both classes by first performing an intra-class study and, afterwards, an inter-class investigation.

Regarding the intra-class variability of the explanations, we were able to demonstrate that the One-Attack framework led to a higher mean distance value for both *bona fide* and *presentation attack* samples. Therefore, it is possible to conclude, from the point of view of interpretability, that the presence of more attacks during training has a positive effect on the generalisation and robustness of the models, confirming our initial intuition. We also examined how the variability of the explanations changed between the two different classes and found that they exhibit similar levels of variability. In addition, we analysed examples of both *bona fide* and *presentation attack* samples, which are representative of the mean variability distance, to get a sense of the intensity of the variability of the explanation from one training setting to another.

This exploratory study confirms the need to establish new approaches in biometrics that incorporate interpretability. There is an urge to evaluate frameworks of explanations that allow the assessment of their quality and comparison. Notably, in the specific use case of PAD, what it means to be a 'good' explanation is not even consensual. For example, a particular region of the face may be relevant for the decision of the model in both cases, whether it is classifying the sample as *bona fide* or *attack*. As pointed out in the literature [41], through the xAI perspective an explanation must describe how the system came to its conclusion, and an 'accurate' explanation (whatever that means) does not imply that a system provided the correct answer. Thus, the open challenge is for the biometrics community to move from the established 'decision accuracy metrics' to novel (not yet developed) 'performance metrics for explanations'. The present work is one step forward in this path by performing a comparison of explanations for the same image using its semantic representation in a space of embeddings where the euclidean distance is an effective similarity metric.

Following the systematic analysis offered in this work, we believe that improving objectivity by combining subjective but knowledgeable opinions from several experts is essential to consolidate interpretability and thus enable deeper and more meaningful performance analysis on presentation attack detection. In addition, further efforts should be made to investigate the impact of the explanations on the training itself, as a regularisation method to guide models through the learning of meaningful features.

ACKNOWLEDGEMENTS

This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia within project UIDB/50014/2020, and within the PhD grants 'SFRH/BD/137720/2018', 'SFRH/BD/139468/2018' and '2020.06434.BD'.

ORCID

Ana F. Sequeira https://orcid.org/0000-0002-6685-2033
Tiago Gonçalves https://orcid.org/0000-0003-4744-9174
Wilson Silva https://orcid.org/0000-0002-4080-9328
João Ribeiro Pinto https://orcid.org/0000-0003-4956-5902
Jaime S. Cardoso https://orcid.org/0000-0002-3760-2473

REFERENCES

1. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature. 521(7553), 436–444 (2015)
2. Karpathy, A., et al.: Large-scale video classification with convolutional neural networks. In: Proceedings of the IEEE CVPR, pp. 1725–1732. Columbus, OH (2014)
3. Samek, W., Wiegand, T., Klaus-Robert, M.: Explainable Artificial Intelligence: Understanding, Visualising and Interpreting Deep Learning Models. Special Issue 1, ICT Discoveries, (2017)
4. Holzinger, A., et al.: Causability and explainabilty of artificial intelligence in medicine. Wiley Interdisc. Reviews: Data Mining and Knowl. Disc., e1312 9(e1312), (2019)
5. Doshi-Velez, F., Kim, B.: Towards a Rigorous Science of Interpretable Machine Learning. arXiv preprint arXiv:1702.08608 (2017)
6. Perez-Cabo, D., et al.: Deep anomaly detection for generalized face anti-spoofing. In: IEEE CVPR Workshops, Long Beach, CA (2019)
7. Bhattacharjee, S., et al.: Recent Advances in Face Presentation Attack Detection. Handbook of Biometric Anti-Spoofing, pp. 207–228. Springer International Publishing, Cham (2019)
8. Rattani, A., Scheirer, W.J., Ross, A.: Open set fingerprint spoof detection across novel fabrication materials. IEEE Trans. Inform. Forensic Secur. 10(11), 2447–2460 (2015)
9. Sequeira, A.F., et al.: A realistic evaluation of iris presentation attack detection. In: 39th TSP, pp. 660–664. Vienna, Austria (2016)
10. Ferreira, P.M., et al.: Adversarial learning for a robust iris presentation attack detection method against unseen attack presentations. In: 2019 International Conference of the Biometrics Special Interest Group (BIOSIG), pp. 1–7. IEEE, Darmstadt, Germany (2019)
11. Sequeira, A.F., et al.: Interpretable biometrics: should we rethink how presentation attack detection is evaluated? In: 2020 8th International Workshop on Biometrics and Forensics (IWBF), pp. 1–6 IEEE, Porto, Portugal (2020)
12. Selvaraju, R.R., et al.: Grad-cam: visual explanations from deep networks via gradient-based localization. In: Proceedings of IEEE ICCV, pp. 618–626. Venice, Italy (2017)
13. ISO/IEC JTC1 SC37: Information technology - biometrics - presentation attack detection Part 3: testing and reporting. ISO Int. Org. Stand. 1, 1–33 (2017)
14. Montavon, G., Samek, W., Müller, K.-R.: Methods for interpreting and understanding deep neural networks. Digit. Signal Process. 73, 1–15 (2018)
15. Kim, B., et al.: Interpretability Beyond Feature Attribution: Quantitative Testing With Concept Activation Vectors (TCAV). 80–2668–2677. Proceedings of the 35th International Conference on Machine Learning preprint arXiv:1711.11279. Stockholm, Sweden (2018)
16. Silva, W., et al.: Towards complementary explanations using deep neural networks. In: Understanding and Interpreting Machine Learning in Medical Image Computing Applications, pp. 133–140. Springer (2018)
17. Silva, W., Fernandes, K., Cardoso, J.S.: How to produce complementary explanations using an ensemble model. In: 2019 International Joint Conference on Neural Networks (IJCNN), pp. 1–8. IEEE, Budapest, Hungary (2019)
18. Samek, W., Müller, K.-R.: Towards Explainable Artificial Intelligence. Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, pp. 5–22. Springer International Publishing (2019)
19. Graziani, M., Andrearczyk, V., Müller, H.: Regression concept vectors for bidirectional explanations in histopathology. In: Understanding and Interpreting Machine Learning in Medical Image Computing Applications, pp. 124–132. Springer (2018)
20. Zhuang, J., et al.: CARE: class attention to regions of lesion for classification on imbalanced data. In: International Conference on Medical Imaging with Deep Learning, pp. 588–597. London, United Kingdom (2019)
21. Arras, L., et al.: What is relevant in a text document?: an interpretable machine learning approach. PloS One. 12(8) (2017)
22. Arras, L., et al.: Explaining Recurrent Neural Network Predictions in Sentiment Analysis. arXiv preprint arXiv:1706.07206 (2017)
23. Bojarski, M., et al.: Explaining How a Deep Neural Network Trained with End-to-End Learning Steers a Car. arXiv preprint arXiv:1704.07911 (2017)
24. Boulkenafet, Z., et al.: A competition on generalized softwarebased face presentation attack detection in mobile scenarios. In: 2017 IEEE International Joint Conference on Biometrics (IJCB), pp. 688–696. IEEE, Denver, CO (2017)
25. Shao, R., et al.: Multi-adversarial discriminative deep domain generalisation for face presentation attack detection. In: Proceedings of the

IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10023–10031. Long Beach, CA (2019)
26. George, A., et al.: Biometric face presentation attack detection with multi-channel convolutional neural network. IEEE Trans. Inf. Forensics Secur. 15, 42–55 (2019)
27. George, A., Marcel, S.: Learning one class representations for face presentation attack detection using multi-channel convolutional neural networks. IEEE Trans. Inf. Forensics Secur. 16, 361–375 (2020)
28. Arashloo, S.R., Kittler, J., Christmas, W.: An anomaly detection approach to face spoofing detection: a new formulation and evaluation protocol. IEEE Access. 5, 13868–13882 (2017)
29. Nikisins, O., et al.: On effectiveness of anomaly detection approaches against unseen presentation attacks in face anti-spoofing. In: 2018 International Conference on Biometrics (ICB), pp. 75–81. IEEE, Gold Coast, Australia (2018)
30. 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, Redondo Beach, CA (2018)
31. Perera, P., Patel, V.M.: Learning deep features for one-class classification. IEEE Trans. Image Process. 28(11), 5450–5463 (2019)
32. Fatemifar, S., et al.: Combining multiple one-class classifiers for anomaly based face spoofing attack detection. In: 2019 International Conference on Biometrics (ICB), pp. 1–7. IEEE, Crete, Greece (2019)
33. Zee, T., Gali, G., Nwogu, I.: Enhancing human face recognition with an interpretable neural network. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. Seoul, South Korea (2019)
34. Liu, Y., Amin, J., 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. Salt Lake City UT, USA (2018)
35. Wang, Z., et al.: Deep spatial gradient and temporal depth learning for face anti-spoofing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5042–5051. Seattle, WA (2020)
36. Yu, Z., et al.: Searching central difference convolutional networks for face anti-spoofing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5295–5305. Seattle, WA (2020)
37. Shao, R., Lan, X., Yuen, P.C.: Regularised fine-grained meta face anti-spoofing. In: Aaai, vol. 34, pp. 11974–11981. New York, NY (2020)
38. Liu, Y., Stehouwer J., Liu, X.: On Disentangling Spoof Trace for Generic Face Anti-Spoofing. arXiv preprint arXiv:2007.09273 (2020)
39. Yang, X., et al.: Face anti-spoofing: model matters, so does data. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3507–3516. Long Beach, CA (2019)
40. Amin, J., Liu, Y., Liu, X.: Face de-spoofing: anti-spoofing via noise modeling. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 290–306. Munich, Germany (2018)
41. Phillips, P.J., Przybocki, M.: Four Principles of Explainable ai as Applied to Biometrics and Facial Forensic Algorithms. arXiv preprint arXiv:2002.01014 (2020)
42. Sharma, R., D-netpad, A.R.: An explainable and interpretable iris presentation attack detector. In: 2020 IEEE International Joint Conference on Biometrics (IJCB), pp. 1–10. IEEE, Houston, TX (2020)
43. Chen, C., Ross, A.: An explainable attention-guided iris presentation attack detector. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops, pp. 97–106. Waikoloa, HI (2021)
44. Seibold, C., Anna, H., Eisert, P.: Focussed lrp: explainable ai for face morphing attack detection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops, pp. 88–96. Waikoloa, HI (2021)
45. Wang, H., et al.: Interpretable security analysis of cancellable biometrics using constrained-optimised similarity-based attack. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops, pp. 70–77. Waikoloa, HI (2021)
46. Joshi, I., et al.: Explainable fingerprint roi segmentation using Monte Carlo dropout. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops, pp. 60–69. Waikoloa, HI (2021)
47. Kaminski, M.E.: The right to explanation, explained. Berkeley Tech. LJ. 34, 189 (2019)
48. Carvalho, D.V., Pereira, E.M., Cardoso, J.S.: Machine learning interpretability: a survey on methods and metrics. Electronics. 8(8), 832 (2019)
49. Hofmanninger, J., Langs, G.: Mapping visual features to semantic profiles for retrieval in medical imaging. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 457–465 (2015)
50. Silva, W., et al.: Interpretability-guided content-based medical image retrieval. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 305–314. Springer, Lima, Peru (2020)
51. Ghorbani, A., et al.: Towards automatic concept-based explanations. In: Advances in Neural Information Processing Systems, pp. 9277–9286. Vancouver, Canada (2019)
52. Zhang, R., et al.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 586–595. Salt Lake City UT, USA (2018)
53. Schroff, F., Kalenichenko, D., James, P.: Facenet: a unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 815–823. Boston, MA (2015)
54. Cao, Q., et al.: Vggface2: a dataset for recognising faces across pose and age. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), pp. 67–74. IEEE, Xi'an, China (2018)
55. Li, H., et al.: Unsupervised domain adaptation for face anti-spoofing. IEEE Trans.Inform.Forensic Secur. 13(7), 1794–1809 (2018)
56. Zhang, K., et al.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process. Lett. 23(10), 1499–1503 (2016)
57. Raghavendra Kotikalapudi and contributors: keras-vis. https://github.com/raghakot/keras-vis (2017)

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Sequeira, A.F., et al.: An exploratory study of interpretability for face presentation attack detection. IET Biome. 10(4), 441–455 (2021). https://doi.org/10.1049/bme2.12045