Myope Models - Are face presentation attack detection models short-sighted?

Pedro C. Neto[1,2], Ana F. Sequeira [1] and Jaime S. Cardoso[2,1]
[1] INESC TEC, Porto, Portugal
[2] Faculdade de Engenharia da Universidade do Porto, Porto, Portugal

Abstract

Presentation attacks are recurrent threats to biometric systems, where impostors attempt to bypass these systems. Humans often use background information as contextual cues for their visual system. Yet, regarding face-based systems, the background is often discarded, since face presentation attack detection (PAD) models are mostly trained with face crops. This work presents a comparative study of face PAD models (including multi-task learning, adversarial training and dynamic frame selection) in two settings: with and without crops. The results show that the performance is consistently better when the background is present in the images. The proposed multi-task methodology beats the state-of-the-art results on the ROSE-Youtu dataset by a large margin with an equal error rate of 0.2%. Furthermore, we analyze the models' predictions with Grad-CAM++ with the aim to investigate to what extent the models focus on background elements that are known to be useful for human inspection. From this analysis we can conclude that the background cues are not relevant across all the attacks. Thus, showing the capability of the model to leverage the background information only when necessary.

1. Introduction

Presentation attacks are one of the weaknesses and dangers posed to face recognition systems (FRS). Apart from a few exceptions [16], most of the methods used in presentation attack detection (PAD) rely on tight face crops [17]. In other words, cropping the faces removes everything in the image that is not part of a face. Thus, the information provided as input for these systems is somehow limited, leading to myopic (i.e. short-sighted) face PAD models. This pre-processing step has several advantages, including the capability to process several faces per frame and it is a fact that first generation face PAD methods were designed to rely only on the face region, in order to be backwards compatible with FRS. However, on the other hand, it removes spatial and contextual information present in the original image. The human visual cortex can process this spatial and contextual information to identify some attacks meant to fool the human inspection. Moreover, in some cases, a human can still be fooled by face crops of replay attacks, if the resolution of the replay device is high enough. Similarly, machine vision systems can likely learn to leverage that information when it is available. Furthermore, they can decide if it is more important to focus on the contextual information or on the face itself.

In this work, we aim at further studying an alternative approach to discarding upfront the background, already adopted in some previous works in the literature. Thus, in this work we investigate how different approaches perform in the presence of contextual background. In the experiments, we perform a comparative study of a state-of-the-art supervised binary classification model and its combination with an adversarial approach (in this, two embeddings are produced, one containing information that is useful for the prediction and the other containing nuisances present in the input, which are optimized to minimize the mutual information between them), on three datasets, using face images with and without crops. For a more thorough evaluation, we elected the ROSE-Youtu dataset due to the fact that it offers a high variety of attacks not available in the other datasets. Thus, we evaluated a multi-task model which was optimized to also distinguish between the different attacks and the previous combined with an adversarial approach. Furthermore, we designed a novel experiment based on multiple instance learning methods. With this, we attempted at creating a dynamic frame selection system, passing the responsibility of selecting the frame most likely to include an attack to the model. Differently from the previous approaches, this method requires more processing power since it goes through all frames in a video.

It has been shown in the literature that black-box models, such as deep neural networks, can learn unpredictable patterns and focus their decision on "unexpected" regions of interest in the input. Therefore, we also evaluated the experiments from the perspective of explainable artificial intelligence (xAI). These evaluations are necessary to better understand the models' decisions and the errors due to their opaqueness [11]. We use methods for the visualization of the elements that were important for the decision, such as Grad-CAM++ [4]. The output of these methods allowed us to analyze the visual cues found by the model to detect attacks. We also note that in some cases, the models follow the same cues used by an human inspector. Thus, we reflect upon the influence of the background in the choice of future face PAD algorithms.

This study contributes to the awareness around the need to incorporate interpretability in face PAD methodologies. The models' performance improvements are attributed to the use of background and this can be corroborated by observing that, in fact, the models' decisions are made using the background cues.

The experiments use three datasets: ROSE-Youtu [20]; NUAA [26]; and Replay-attack [5]. The state-of-the-art supervised binary classification model (BC), and BC combined with an adversarial approach are evaluated on the three datasets in two settings: with and without background. Furthermore, we evaluate the multi-task (MT) and dynamic frame selection (DFS) approaches using only the ROSE-Youtu dataset which includes a high diversity of attacks comprising both two-dimensional and three-dimensional information. Hence, it was used to study the impact of the background in the model performance and whether the background affects the capability of generalizing between attacks. However, the experiments were defined with two goals in mind: generalization between attacks, and generalization between subjects. The first goal is addressed, through the study of the performance of the model on attacks that were not seen previously during training. The second is addressed by using 50% of the subjects for testing. The BC, MT and adversarial approaches are evaluated with and without background in a cross-dataset scenario.

The major contributions of this work are: i) the evaluation (in three widely used datasets) of a state-of-the-art supervised binary classification model and its combination with an adversarial strategy in two alternative scenarios: face images with and without background; ii) a multi-task face PAD approach that leverages background and achieves state-of-the-art results on the ROSE-Youtu dataset; iii) a proposed methodology for frame selection strategy on the ROSE-Youtu dataset. It was not the focus of this study to see if specific models in the literature perform better with and without background, instead, it focuses on proposing simple and distinct approaches and analyzing whether the results are consistent across them with regard to the presence of background.

Besides this introduction and the conclusion, this document contains four major sections. First, in Section 2 there is a discussion of the related work and how it led to the current study. Afterwards, details on the experiments conducted are given in Section 3. The description of the datasets is given in Section 4. And finally, in Section 5 we present and discuss the obtained results and how it impacts the future of PAD methods.

2. Related Work

Typically, previous works on face presentation attack detection do not leverage background information. And thus, removing it is a common practice and a frequent step on the preprocessing stage. The most common approach is to use a face crop, usually obtained through the use of deep learning-based face detection algorithms such as MTCNN [32] and RetinaFace [6]. Earlier methods used more traditional techniques, for instance, the Viola-Jones cascade detector [28].

Within the published works, it is possible to find reinforcement learning approaches [3], 3D-CNNs [19], a two stages approach relying on blinking [8] and several other colour-based methods [20, 2]. The background usage is addressed in some works [27, 1, 22, 30, 16], however, they did not perform comparative studies regarding performance, with and without the background, of several approaches. It is possible to find this comparison in other works, however, the proposed methods are based on conventional machine learning instead of end-to-end deep learning [31, 18]. Due to the nature of the ROSE-Youtu dataset, which contains three-dimensional and two-dimensional attacks, there are fewer methods tested on this dataset than on others. The variability of attacks included in the dataset significantly increase the difficulty of finding a model capable of performing well on all of them. For this reason, even methods that achieve almost zero error on other datasets, have worse performance on the ROSE-Youtu [7].

To the best of our knowledge, there has not been any method inspired by multiple instance learning applied to face PAD. However, there is an article on a similar technique used for the detection of deep fakes [21]. Despite being a slightly different problem, the detection methodology has a significant overlap. The adversarial approach followed in our experiments was described first at [13] and its capabilities to work with face PAD systems was evaluated one year later by Jaiswal *et al.* [14].

Producing and visualizing explanations of the predictions for face presentation attack detection is a relatively new topic. Sequeira *et al.* have explored the challenge of interpreting face PAD methods [25, 24]. Their work described how the current evaluation metrics for PAD lack information regarding the elements that are being used for the prediction. In a sense, they argue that models can make accurate predictions but still base their decision on parts of the image that do not correspond to real face features or presentation artifacts as a human inspector would. In this work, we follow a similar approach to produce and analyze explanations. However, we use them to infer if the presence of contextual background leads to the use of certain visual cues in the image. At the same time, we look forward to seeing if other contextual elements are used to make correct predictions, for instance, reflections. Humans often use these elements to make their analysis.

3. Methodology

Myriads of attacks are constantly threatening biometric systems. However, in practice, we do not aim to identify the type of attack, thus the main goal is to infer if the image given to the sensor is an attack or if it is from a genuine person. The problem is, in its essence, formulated as a binary classification task. On top of the binary task, we also applied some different training processes. However, these do not affect the network at test time. The purpose of these distinct approaches is to understand if the background effect generalizes between approaches.

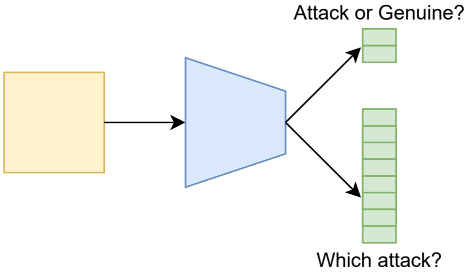


Figure 1. Architecture of the multi-task learning model. It receives an image (yellow box) and includes a CNN (blue figure), two output heads (green figures), where the first is a binary head and the second has 8 output nodes. The output node for the genuine samples on both heads ensures that the main goal remains on the detection of attacks.

- *Binary classification training (BC)* - In the first approach, the only task that the backbone network is optimized for at training time is to classify between attacks and genuine. For this, we use a MobileNet v2 [23] that outputs two values, which are activated with softmax. The optimization of the weights is done using the binary cross-entropy loss (Eq. 1).

$$BCE(y,p) = -(y\log(p) + (1-y)\log(1-p)) \tag{1}$$

- *Multi-task classification training (MT)* - Whenever a model is optimized to distinguish between attacks and genuine images, it treats all the attacks equally. However, in practice, the attacks are not the same, and each possesses distinctive characteristics. And thus, we also formulated the training stage of a MobileNet v2, so it learns to distinguish between the seven different attacks. It is also possible that learning to discriminate between attacks also boosts the performance whenever the attack is unknown. Instead of having an output layer with two classes, the network has two output layers. The first has two output classes, whereas the other has eight (seven attacks and one genuine). In both cases, they are activated with softmax. Both layers are, simultaneously, updated with the binary cross-entropy 1 and cross-entropy 2 losses, respectively. These losses are combined as seen on Equation 3. Due to the risk of the second term of the equation being larger than the first, it was necessary to add an output node for genuine samples in both heads. Figure 1 is a simplified visualization of the architecture of the model.

$$CE(y,p) = -\sum_{c=1}^{M} y_{o,c}\log(p_{o,c}) \tag{2}$$

$$Loss_{Multi}(y_1,p_1,y_2,p_2) = BCE(y_1,p_1) + CE(y_2,p_2) \tag{3}$$

- *Adversarial training (Adv.)* - In the images shown to the system, there is background information that is useful for the prediction, for instance, reflections. Nevertheless, not all the background information is useful. Part of it can be considered to be a nuisance. Hence, we explored also an approach that attempted to remove those parts of the image from the feature vector used for the classification task. This approach, known as Unsupervised Adversarial Invariance [13], produces two distinct embeddings (i.e. feature vectors). The first vector, e_1, represents the features that are relevant to the prediction of the model, whereas the second, e_2, comprises the information that should not be used for the prediction. Constructing the loss of such architecture requires four terms. The first two are maximization terms (Eq. 6 and 7). They attempt to reconstruct e_1 from e_2 and vice-versa. This attempts at removing any potential mutual information between both embeddings. The other two are minimization terms (Eq. 5). The first embedding uses e_1 to perform the classification task. Whereas the second apply some noise to e_1, in the form of a dropout layer. From the noisy e_1 and from e_2, it tries to reconstruct the input image. For the construction/reconstruction terms the loss used is the mean squared error (Eq. 4), while for the classification term we use either the binary cross-entropy (Eq. 1) or the multi-task loss (Eq. 3). The term α controls the impact of the reconstruction loss on the overall loss. We start with $\alpha = 0.025$, and we increase by 0.025 at the end of each epoch. The architecture is represented in Figure 2.

$$MSE(x,y) = \sum_{i=1}^{D}(x_i - y_i)^2 \tag{4}$$

$$Loss_{Adv}(e_1, e_2, e'_1, e'_2) = -MSE(e_1, e'_2) - MSE(e_2, e'_1) \tag{5}$$

$$Loss_{Class}(y, p, x, x') = BCE(y, p) + \alpha MSE(x, x') \tag{6}$$

$$\begin{aligned} Loss_{Class}(y_1, p_1, y_2, p_2, x, x') = \\ Loss_{Multi\_task}(y_1, p_1, y_2, p_2) \\ + \alpha MSE(x, x') \end{aligned} \tag{7}$$

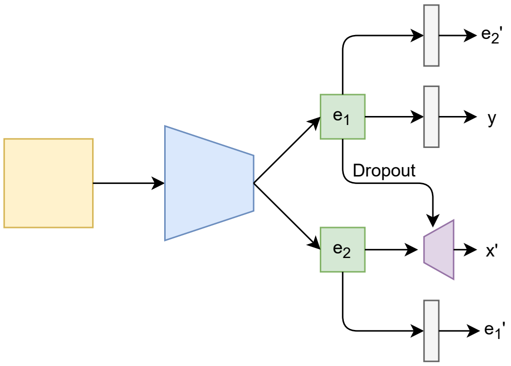


Figure 2. Architecture of the adversarial learning model. It receives an input image (yellow box) and includes a CNN (blue figure), two feature vectors e_1 and e_2 (green boxes). These are used to reconstruct each other, decode (purple figure) the input and to classify the input. This architecture is deeply based on the Unsupervised Adversarial Invariance [13].

- *Dynamic frame selection training (DFS)* - Finally, we also propose an architecture to select the best frame for the detection of attacks. In state-of-the-art approaches, the training frames were fixed and previously selected from the list of possible frames. This, however, raises a couple of questions: 1) Do all the frames contain the same information for the prediction?; 2) If not, are we selecting the best frames to optimize the network?. We structured the optimization of this method in two stages: frame selection and learning. The frame selection stage processes all the frames in a video and computes their output with the model. From the outputs, if the video is from an attack it selects the three frames that have the lowest probability of being an attack. And the opposite if the video is from a genuine individual. Afterwards, the selected frames are used in the learning stage to optimize the network towards the video labels. At testing time the process is similar to the frame selection, the frame with the highest probability of being an attack is used for the classification. Perhaps, one of the most interesting aspects of this approach is that it can be integrated with the other previously mentioned. The frame selection step remains unchanged, while the training stage integrates the changes related to the other approaches. Figure 3 shows the behavior of the described method for both frame selection and testing. The frame selection probability is the result of the attack probability produced by the binary classification layer of the model.

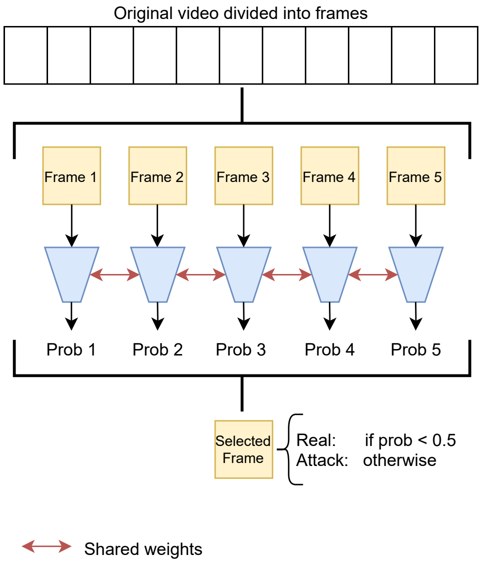


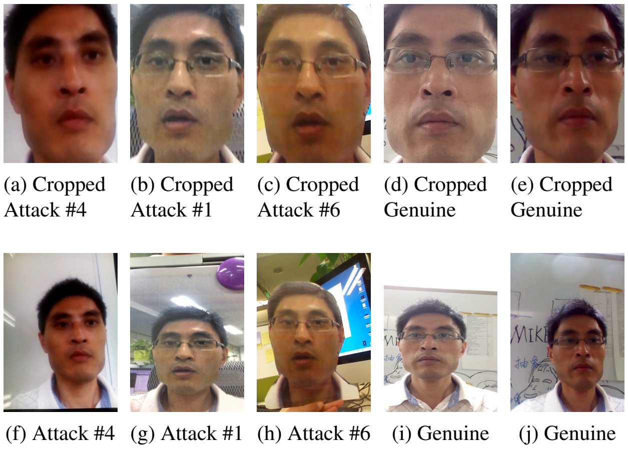
Figure 3. Architecture of the method for dynamic frame selection. It includes a CNN (blue figure) with shared weights that processes all the frames (yellow boxes) of a video and selects one (lower yellow box) based on a specific criteria.

To evaluate and compare the performance of these PAD models, we collected the following metrics: 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) [12]. Finally, we also collected the *Equal Error Rate (EER)*, which is the error at the operation point where the APCER and BPCER have the same value. For the APCER and the BPCER we used a threshold of 0.5.

4. Datasets

The datasets used for the experimental evaluation are: ROSE-Youtu [20]; NUAA [26]; and Replay-Attack [5].

ROSE-Youtu [20]: Contains, in its public version, 3350 videos with 20 different subjects. On average, video clips have a duration of 10 seconds. For each of the subjects, it contains around 150 to 200 videos captured from five mobile devices (all with different resolutions on their camera) and five lighting conditions. The front-facing camera was used with a distance between face and camera of about 30 to 50 centimeters.


(a) Cropped Attack #4 (b) Cropped Attack #1 (c) Cropped Attack #6 (d) Cropped Genuine (e) Cropped Genuine

(f) Attack #4 (g) Attack #1 (h) Attack #6 (i) Genuine (j) Genuine

Figure 4. Samples collected from the ROSE-YOUTU dataset [20] containing images from attacks and genuine captures. On the top row, cropped images are displayed. Whereas the bottom row contains the exact same images, but with all the background information included.

Table 1. List of attacks present in the ROSE-YOUTU dataset [20].

Attack	Description
-	Genuine (bona fide)
#1	Still printed paper
#2	Quivering printed paper
#3	Video which records a Lenovo LCD display
#4	Video which records a Mac LCD display
#5	Paper mask with two eyes and mouth cropped out
#6	Paper mask without cropping
#7	Paper mask with the upper part cut in the middle

There are eight different types of videos, which translates into eight classes. The first class represents the genuine samples, whereas each of the following seven represent an attack. The first two attacks are print attacks, while the third and fourth are replay attacks on a Lenovo and an Apple laptop, respectively. The remaining three are based on paper masks and are responsible for including three-dimensional information in the dataset. These attacks are described in Table 1.

We preprocessed the dataset into two different copies. For both, the frames of the videos were extracted and stored. For the second, a face was cropped by the MTCNN algorithm [32] for all frames. Examples of these images are seen in Figure 4. We used the videos from the first 10 indexed subjects (2,3,4,5,6,7,9,10,11,12) for training and the remaining 10 for testing.

NUAA [26]: was one of the first public databases for training and evaluating the performance of face PAD methods. This database simulates a simple and general method that re-captures a printed photograph of users for attacking a face recognition system. The NUAA database contains real and presentation attack face images of 15 persons. For each person, both real and presentation attack images were captured in three different sessions using generic cheap webcams and real face and printed photograph of users. The NUAA database contains 5105 real and 7509 presentation attack face images in color space with 640 × 480 pixels of image resolution. In this database, using the collected images, the training and testing sub-databases are predefined for training and testing of the PAD method, through which the performances of various PAD methods can be compared. In detail, the training database contains 1743 real and 1748 presentation attack face images, while the testing database contains 3362 real and 5761 presentation attack face images.

Replay-Attack [5]: this database for face PAD consists of 1300 video clips of photo and video attack attempts to 50 clients, under different lighting conditions. All videos are generated by either having a (real) client trying to access a laptop through a built-in webcam or by displaying a photo or a video recording of the same client for at least 9 seconds.

5. Results

In this section, we present and discuss the evaluation results of the methods described. Regarding implementation details, all the methods described were optimized with Adam and a fixed learning rate of 0.001. The model used is a MobileNet v2 pre-trained on the ImageNet dataset. The images are resized to have a resolution of 224x224 and an RGB color scheme. The Grad-CAM++ [4] was the visualization tool used to analyze the parts of the image relevant to the models' decisions. Train and test set splits were the same described in the publication of each dataset, so that the results can be compared with other works.

In Table 2 are shown the results obtained with the binary classification (BC) and its combination with the adversarial training (Adv.+BC) for the NUAA and Replay-Attack datasets. From both evaluations it is evident the performance improvement when the background is present in the images with a decrease in the EER to 0.00%.

The variety of attacks present in the ROSE-Youtu dataset motivates a multi-task learning approach. The experiments produced intended to evaluate the BC and the MT approaches with the inclusion and exclusion of contextual background. We further attempted to integrate them with the adversarial approach described in the previous section. The results for both BC and MT classification and their combination with the adversarial training strategy (Adv.+BC; Adv.+MT) can be seen in Table 3. While the adversarial training did not lead to the expected improvement, in all four scenarios the models' performance improved with the background. This seems to indicate that the background provides more information and favoured the performance error rates. On multi-task classification, the improvements on the EER are as high as 81%.

Table 2. Comparison of four different approaches with their versions with and without background. The columns represent the dataset used for both training and testing. The reported values represent the EER in %.

Method	Background	NUAA	Replay-Attack
BC	No	2.91	**0.00**
	Yes	**0.00**	**0.00**
Adv.+BC	No	3.03	0.33
	Yes	**0.00**	**0.00**

Table 3. Comparison of four different approaches with their versions with and without background on the ROSE-Youtu dataset. APCER, BPCER and EER are displayed as %. In bold is the best result per column.

Method	Background	APCER	BPCER	EER
BC	No	0.49	2.20	1.32
	Yes	**0.25**	**2.03**	**0.73**
MT	No	1.34	1.17	1.26
	Yes	**0.15**	**0.40**	**0.24**
Adv.+BC	No	1.42	2.71	1.76
	Yes	**0.52**	**1.29**	**0.76**
Adv.+MT	No	1.18	2.93	1.91
	Yes	**0.29**	**1.11**	**0.60**

In a cross-dataset approach, we performed experiments in which the models trained with the ROSE-Youtu dataset were evaluated with the other two datasets. The results of these experiments are presented in Table 4. These results are in line with the results depicted on Tables 2 and 3. The comparison of the same-database and cross-database results show that the models' performance consistently improve when the background information is used.

Considering the performance gains from the use of background, we used this scenario to explore the proposed multi-task (MT) and dynamic frame selection (DFS) strategies. The results of all approaches evaluated for the ROSE-Youtu dataset are presented in Table 5. Unexpectedly, the performance of the DFS methods and the ones that used adversarial training produced worse results than the simple binary and multi-task classification. The BC and MT approaches performed well at detecting attacks, as can be seen by the low value of the APCER, 0.25% and 0.15%, respectively. Regarding the detection of bonafide images, the BC performed worse than several of the other methods and the MT had the lowest BPCER of them all, 0.40%.

Table 4. Comparison of four different approaches with their versions with and without background. Results for models trained on the ROSE-Youtu dataset and tested on the datasets of each column. The reported values represent the EER in %.

Method	Background	NUAA	Replay-Attack
BC	No	22.04	29.43
	Yes	**13.45**	**12.29**
MT	No	23.61	26.13
	Yes	**3.89**	**13.91**
Adv.+BC	No	28.31	26.03
	Yes	**18.11**	**17.12**
Adv.+MT	No	35.66	26.15
	Yes	**23.85**	**19.46**

Table 5. Comparison of all the seven different approaches explored in the ROSE-Youtu dataset. All of the approaches leveraged background information. In bold is the best result per column.

Method	APCER (%)	BPCER (%)	EER (%)
BC	0.25	2.03	0.73
MT	**0.15**	**0.40**	**0.24**
Adv. + BC	0.52	1.29	0.76
Adv. + MT	0.29	1.11	0.60
DFS	0.54	4.68	1.62
MT + DFS	0.31	2.23	0.69
Adv. + DFS	2.15	1.78	1.78

We extended the experiments of the multi-task classification approach for different evaluation configurations, one and unseen attack (the results can be seen in Table 6 and Table 7, respectively). The multi-task classification was also integrated with the adversarial training and the dynamic frame selection for a better and more complete comparison.

The one attack configuration selects one attack to be used for both training and testing. This is intended to see how hard is to overfit the model to that attack and to distinguish it from genuine images. The results for this configuration are visible in Table 6 and it is possible to see that while the three approaches are capable of overfitting to the majority of the attacks, the attack #4 remains challenging to the adversarial and DFS approaches.

The unseen configuration selects one attack to be removed from training and to be the only one used for testing. This is to evaluate the capability of the model to generalize to novel attacks and to evaluate the challenges that each attack present to the network. The results for this configuration are seen in Table 7, and it is possible to observe that both DFS and adversarial approaches have difficulties in generalizing to unseen attacks. Especially attack #4. The low performance of this attack can be explained by the high resolution of the replay attack device that increases the difficulty of the task.

Table 6. Evaluation of three approaches in the setting of one attack in the ROSE-Youtu dataset. In this setting, the attack in the first column is the only one used for training and testing. APCER, BPCER and EER are displayed as %. In bold is the best result per column.

Attack	MT			Adversarial MT			DFS MT		
	APCER	BPCER	EER	APCER	BPCER	EER	APCER	BPCER	EER
#1	**0.00**	**0.20**	**0.05**	0.20	0.29	0.27	0.50	0.22	0.50
#2	**0.00**	0.02	0.02	**0.00**	0.07	**0.00**	**0.00**	**0.00**	**0.00**
#3	**0.00**	**0.16**	0.11	**0.00**	0.29	0.25	**0.00**	0.45	**0.00**
#4	**0.30**	**1.27**	**0.71**	1.16	1.98	1.46	0.50	1.34	1.01
#5	**0.00**	**0.02**	**0.00**	**0.00**	0.09	0.05	**0.00**	0.45	**0.00**
#6	**0.00**	0.07	**0.00**	**0.00**	0.16	0.11	**0.00**	**0.00**	**0.00**
#7	**0.00**	0.05	**0.00**	**0.00**	0.11	**0.00**	**0.00**	**0.00**	**0.00**

Table 7. Evaluation of three approaches in the setting of unseen attack in the ROSE-Youtu dataset. In this setting, the attack in the first column is excluded from the training and is the only one used for testing. APCER, BPCER and EER are displayed as %. In bold is the best result per column.

Attack	MT			Adversarial MT			DFS MT		
	APCER	BPCER	EER	APCER	BPCER	EER	APCER	BPCER	EER
#1	**1.00**	**0.85**	**0.95**	**1.00**	6.90	2.87	**1.00**	6.90	2.30
#2	**0.00**	**0.49**	**0.25**	**0.00**	1.18	0.27	**0.00**	4.90	0.67
#3	3.09	3.41	3.27	**1.44**	**3.07**	**2.61**	2.14	10.24	5.79
#4	13.82	**3.88**	**7.57**	**12.66**	6.61	9.15	17.59	16.70	17.09
#5	**0.00**	1.98	**0.65**	0.25	**0.58**	**0.45**	0.50	4.90	1.49
#6	**0.00**	0.89	0.33	0.10	**0.58**	**0.10**	**0.00**	4.01	1.00
#7	**0.35**	**3.63**	**1.77**	2.32	7.68	5.20	0.51	9.58	1.78

The multi-task approach excelled at detecting both attacks and bonafide samples, achieving an equal-error rate better than any other approach. And thus, it was the approach used to compare with the state-of-the-art for the ROSE-YOUTU dataset. he methods compared report their results on similar train/test split, as specified on the database publication document [20]. Our results, as seen in Table 8 are better than the state-of-the-art when we include background and slightly better when there is no background. Despite the good results, it is important to note that the methods presented in this document were not designed to be the best performing methods at cross-dataset configuration. Instead, they are intended to allow a relevant study regarding the presence of background for this specific dataset.

Table 8. Comparison of the best proposed approaches, both with and without background, with the state-of-the-art. In bold is the best result per column.

Method	EER (%)
CoALBP (YCBCR) [20]	17.1
CoALBP (HSV) [20]	16.4
Color [2, 7]	13.9
De Spoofing [15, 7]	12.3
RCTR-all spaces [7]	10.7
ResNet-18 [9]	9.3
SE-ResNet18 [10]	8.6
AlexNet [20]	8.0
DR-UDA (SE-ResNet18) [29]	8.0
DR-UDA (ResNet-18) [29]	7.2
3D-CNN [19]	7.0
Blink-CNN [8]	4.6
DRL-FAS [3]	1.8
Ours w/ Background	**0.2**

In Figure 5 is depicted the ROC curve of the MT model with the x-axis displayed at the log-scale for a better visualization. The model is indeed nearly perfect at detecting the attacks and the bona-fide images in the ROSE-Youtu dataset.

Finally, we produced explanations of our model for an example of each category of attacks. For the replay attack, we produced the explanations in Figures 6a and 6d. In these figures, it is possible to observe that the models leveraged the presence of reflections in the attack image, whenever there is background. When the background is not present there are no cues to justify the decision of the model, which is in fact wrong. Figures 6b and 6e shows the explanations for a paper mask attack, and as expected, the explanations do not rely on the background. Instead, the model directs its focus to the mask area for the final prediction. The area is similar on both versions of the model with and without

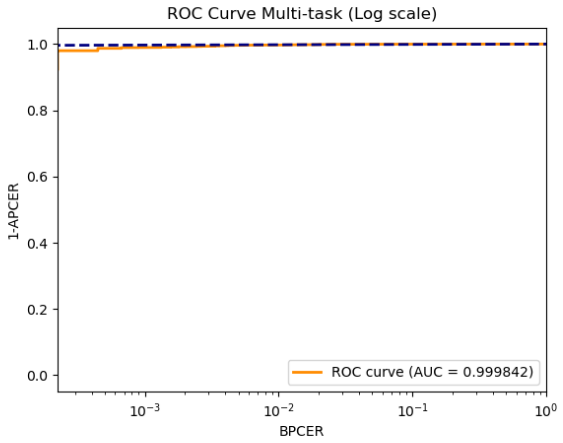
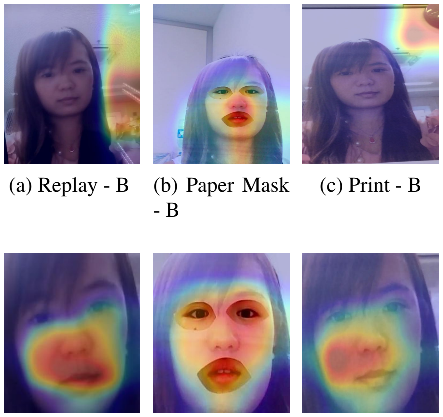


Figure 5. Receiver operating characteristic curve for the multi-task model on the ROSE-Youtu dataset with background. X-axis is at log-scale.



(a) Replay - B (b) Paper Mask - B (c) Print - B

(d) Replay (e) Paper Mask (f) Print

Figure 6. Explanations generated with Grad-CAM++ for different attacks of subject #23. -B indicates that the model was trained and tested on images with background. The (d) image was wrongly classified as genuine.

background. Finally, the print attack explanations are seen in Figures 6c and 6f. These figures shows that once more the model is capable of understanding the conditions of the image given and directs its focus to an important background artefact, the pin holding the image. And again, the version without background does not highlight any relevant cue that explains the prediction of the model. Hence, we further argue in favor of a better explanability factor in the models that include background.

6. Conclusion

This work explored how consistently the background impacts the performance of distinct methods for face presentation attack detection. The experiments corroborated the view that a face PAD model is capable of leveraging both background and face elements to make a correct prediction.

Our approach surpassed the state-of-the-art results for the ROSE-YOUTU dataset by a large margin. The multi-task model leverages background artefacts to improve the detection of specific attacks. Moreover, we also present some alternative approaches, dynamic frame selection and adversarial training, that we believe were limited by the lack of a large database of face presentation attacks. Their results were consistent with the one from the multi-task model.

We further contribute to improve the explainability of these models by analyzing the predictions. This analyze conducted with the Grad-CAM++ algorithm highlighted that models that include the background of the images can leverage the presence of certain artifacts. On the other hand, when the background is not present the generated explanations seem to be non-informative. Hence, due to their similarity with the human vision with regards to the areas used for the prediction, models that leverage the background provide more explanations for their predictions.

Acknowledgments

The authors would like to acknowledge the reviewers' comments that were crucial for the improvement of this work. This work was 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 grant “2021.06872.BD”.

Conflict of Interest

The authors declare that there is no conflict of interest that could be perceived as prejudicing the impartiality of the research reported.

References

[1] Yashasvi Baweja, Poojan Oza, Pramuditha Perera, and Vishal M Patel. Anomaly detection-based unknown face presentation attack detection. In *2020 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–9. IEEE, 2020.

[2] Zinelabidine Boulkenafet, Jukka Komulainen, and Abdenour Hadid. Face spoofing detection using colour texture analysis. *IEEE Transactions on Information Forensics and Security*, 11(8):1818–1830, 2016.

[3] Rizhao Cai, Haoliang Li, Shiqi Wang, Changsheng Chen, and Alex C Kot. Drl-fas: A novel framework based on deep reinforcement learning for face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 16:937–951, 2020.

[4] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 839–847, 2018.

[5] Ivana Chingovska, André Anjos, and Sébastien Marcel. On the effectiveness of local binary patterns in face anti-spoofing. In *2012 BIOSIG-proceedings of the international conference of biometrics special interest group (BIOSIG)*, pages 1–7. IEEE, 2012.

[6] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[7] Yuting Du, Tong Qiao, Ming Xu, and Ning Zheng. Towards Face Presentation Attack Detection Based on Residual Color Texture Representation. *Security and Communication Networks*, 2021:6652727, 2021.

[8] Md. Mehedi Hasan, Md. Salah Uddin Yusuf, Tanbin Islam Rohan, and Shidhartho Roy. Efficient two stage approach to detect face liveness : Motion based and deep learning based. In *2019 4th International Conference on Electrical Information and Communication Technology (EICT)*, pages 1–6, 2019.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[10] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[11] Matthew Hutson. The opacity of artificial intelligence makes it hard to tell when decision-making is biased. *IEEE Spectrum*, 58(2):40–45, 2021.

[12] ISO/IEC JTC1 SC37. Information Technology - Biometrics - Presentation attack detection Part 3: Testing and Reporting. *ISO Int. Organization for Standardization*, 2017.

[13] Ayush Jaiswal, Rex Yue Wu, Wael Abd-Almageed, and Prem Natarajan. Unsupervised Adversarial Invariance. In S Bengio, H Wallach, H Larochelle, K Grauman, N Cesa-Bianchi, and R Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[14] Ayush Jaiswal, Shuai Xia, Iacopo Masi, and Wael AbdAlmageed. Ropad: Robust presentation attack detection through unsupervised adversarial invariance. In *2019 International Conference on Biometrics (ICB)*, pages 1–8. IEEE, 2019.

[15] Amin Jourabloo, Yaojie Liu, and Xiaoming Liu. Face de-spoofing: Anti-spoofing via noise modeling. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 290–306, 2018.

[16] Alperen Kantarcı, Hasan Dertli, and Hazım Kemal Ekenel. Shuffled patch-wise supervision for presentation attack detection. In *2021 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–5. IEEE, 2021.

[17] Dakshina Ranjan Kisku and Rinku Datta Rakshit. Face spoofing and counter-spoofing: a survey of state-of-the-art algorithms. *Transactions on Machine Learning and Artificial Intelligence*, 5(2):31, 2017.

[18] Jukka Komulainen, Abdenour Hadid, and Matti Pietikäinen. Context based face anti-spoofing. In *2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, pages 1–8. IEEE, 2013.

[19] Haoliang Li, Peisong He, Shiqi Wang, Anderson Rocha, Xinghao Jiang, and Alex C. Kot. Learning generalized deep feature representation for face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 13(10):2639–2652, 2018.

[20] Haoliang Li, Wen Li, Hong Cao, Shiqi Wang, Feiyue Huang, and Alex C. Kot. Unsupervised domain adaptation for face anti-spoofing. *IEEE Transactions on Information Forensics and Security*, 13(7):1794–1809, 2018.

[21] Xiaodan Li, Yining Lang, Yuefeng Chen, Xiaofeng Mao, Yuan He, Shuhui Wang, Hui Xue, and Quan Lu. Sharp multiple instance learning for deepfake video detection. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1864–1872, 2020.

[22] Ewa Magdalena Nowara, Ashutosh Sabharwal, and Ashok Veeraraghavan. Ppgsecure: Biometric presentation attack detection using photopletysmograms. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 56–62. IEEE, 2017.

[23] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[24] Ana F. Sequeira, Tiago Gonçalves, Wilson Silva, João Ribeiro Pinto, and Jaime S. Cardoso. An exploratory study of interpretability for face presentation attack detection. *IET Biometrics*, n/a(n/a).

[25] A. F. Sequeira, W. Silva, J. R. Pinto, T. Gonçalves, and J. S. Cardoso. Interpretable biometrics: Should we rethink how presentation attack detection is evaluated? In *2020 8th International Workshop on Biometrics and Forensics (IWBF)*, pages 1–6, 2020.

[26] Xiaoyang Tan, Yi Li, Jun Liu, and Lin Jiang. Face liveness detection from a single image with sparse low rank bilinear discriminative model. In *European Conference on Computer Vision*, pages 504–517. Springer, 2010.

[27] Rafael Henrique Vareto and William Robson Schwartz. Face spoofing detection via ensemble of classifiers toward low-power devices. *Pattern Analysis and Applications*, 24(2):511–521, 2021.

[28] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–I, 2001.

[29] Guoqing Wang, Hu Han, Shiguang Shan, and Xilin Chen. Unsupervised adversarial domain adaptation for cross-domain face presentation attack detection. *IEEE Transactions on Information Forensics and Security*, 16:56–69, 2021.

[30] Jianwei Yang, Zhen Lei, and Stan Z Li. Learn convolutional neural network for face anti-spoofing. *arXiv preprint arXiv:1408.5601*, 2014.

[31] Jianwei Yang, Zhen Lei, Shengcai Liao, and Stan Z Li. Face liveness detection with component dependent descriptor. In

2013 International Conference on Biometrics (ICB), pages 1–6. IEEE, 2013.

[32] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.