Pre-print of article that will appear in IEEE Transactions on Information Forensics and Security, vol. 15, n. 1, pp. 3347-3358, 2020, doi: 10.1109/TIFS.2020.2988168.

The published article is available on https://ieeexplore.ieee.org/document/9068268/

Leveraging Shape, Reflectance and Albedo from Shading for Face Presentation Attack Detection

Allan Pinto, *Member, IEEE,* Siome Goldenstein,
Alexandre Ferreira, Tiago Carvalho *Member, IEEE*, Helio Pedrini, *Senior Member, IEEE*,
and Anderson Rocha, *Senior Member, IEEE*

***Abstract*—Presentation attack detection is a challenging problem that aims at exposing an impostor user seeking to deceive the authentication system. In facial biometrics systems, this kind of attack is performed using a photograph, video, or 3D mask containing the biometric information of a genuine identity. In this paper, we propose a novel approach to detecting face presentation attacks based on intrinsic properties of the scene such as albedo, depth, and reflectance properties of the facial surfaces, which were recovered through a shape-from-shading (SfS) algorithm. To extract meaningful patterns from the different maps obtained with the SfS algorithm, we designed a novel shallow CNN architecture for learning features useful to the presentation attack detection (PAD). We performed several experiments considering the intra- and inter-dataset evaluation protocols. The obtained results showed the effectiveness of the proposed method considering several types of photo- and video-based presentation attacks, and in the cross-sensor scenario, besides achieving competitive results for the inter-dataset evaluation protocol.**

***Index Terms*—Face Presentation Attack Detection, Face Spoofing Attack Detection, Facial Biometric System, Shape-from-Shading, Albedo, Reflectance, Depth, Intrinsic Properties of Surface, Surface Reconstruction, Convolutional Neural Network, Deep Learning.**

I. Introduction

BIOMETRICS is an active research field whose today's challenges go far beyond obtaining a high precision system. Nowadays, security aspects of biometric systems are essential for a successful authentication mechanism due to the vast possibility that an impostor user has for attacking it. Among these possibilities, a presentation attack is the easiest way to deceive such systems. This kind of attack can be performed directly on the acquisition sensor without any previous knowledge of the internal components of the system. It is characterized by the action of presenting a synthetic biometric sample, such as photographs, digital video, or even a 3D mask, of a valid user to the acquisition sensor in order to authenticate itself as a legitimate user [1].

Although several advances have been reported in the literature, face presentation attack detection (PAD) is still an open problem. According to the Intl. Joint Conference on Biometrics (IJCB) 2017 competition on generalized face presentation attack detection in mobile devices [2], the best algorithm for detecting presentation attacks (PA) presented an Attack Presentation Classification Error Rate (APCER) of 5.0% under environmental, attack types, and camera device variations. Thus, five out of one every hundred attempted attacks were successfully accomplished, which may render the authentication process unfeasible in practice if we consider a system with hundreds of thousands of users.

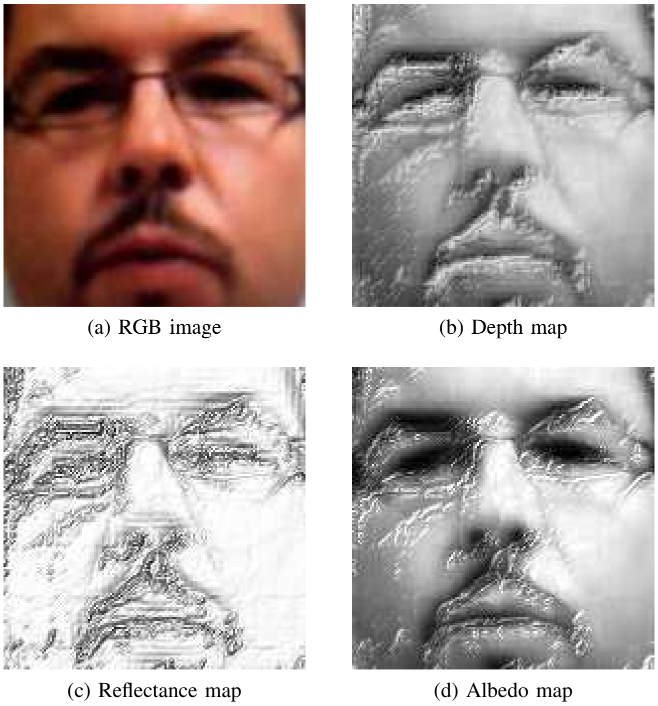

(a) RGB image (b) Depth map
(c) Reflectance map (d) Albedo map

Fig. 1. Example of a facial surface reconstruction using an SfS algorithm for presentation attack video frame.

Currently, a major limitation of existing solutions for PAD is the lack of ability to work in an unknown environment. In fact, PAD solutions available in the literature present impressive accuracy rates, with near-perfect classification results, when they are trained and tested with data from same source. However, when we consider challenging evaluation scenarios, those algorithms present low performance, sometimes becoming worse than random. For this reason, researchers have promoted efforts to report their results considering two evaluation protocols known as intra- and inter-datasets. An intra-dataset evaluation protocol consists of testing a PAD

A. Pinto, A. Ferreira, T. Carvalho, H. Pedrini and A. Rocha are with the Institute of Computing, University of Campinas (Unicamp), Av. Albert Einstein, 1251, Campinas, SP, Brazil, 13083-852. E-mail: {allan.pinto,helio,anderson.rocha}@ic.unicamp.br.

T. Carvalho is also with Federal Institute of São Paulo (IFSP), Campinas, Brazi

S. Goldenstein is with Google Inc., Pittsburgh, PA, USA. E-mail: siome@google.com.

Manuscript received ...; revised ....

algorithm using data that came from the same source as the training data, whose samples were collected using the same acquisition sensor and in the same environment. In turn, an inter-dataset protocol consists of testing a PAD algorithm using data from a different source as the training data, which means we have data from different domains (i.e., different sensors and environments). Such evaluation protocol is more challenging and more suitable for reflecting a real operating scenario.

According to recent results reported in the literature, the Half Total Error Rates (HTER) can increase drastically taking into account inter-dataset evaluation protocol [3]. It is the case of Pinto *et al.* [4] work that proposed a PAD method based on analysis of the noise and artifacts left in the synthetic biometric sample during its manufacture such as blurring, printing effect, banding effect among others. Although the authors achieved a low HTER value for the intra-dataset evaluation protocol (2.8%), the HTER of this technique increases significantly, considering the inter-dataset protocol (34.4%). Even in the state-of-the-art techniques, values for error rates are still too high. Boulkenafet *et al.* [5] reported an HTER of 2.9% and 16.7% considering the intra- and inter-dataset protocols, respectively, which means a relative change in terms of HTER of about 475%, which is far from an acceptable value in practice.

In this paper, we present a novel approach to distinguish a synthetic face from real ones, taking into account optical and physical properties of the scene captured by the acquisition sensor. Our method takes advantage of the depth information, associating it with light properties of the scene to detect an attempted attack, using a technique known as shape-from-shading (SfS). SfS was firstly proposed by Horn *et al.* [6] and aims to estimate the shape of an object based on the shade information present in its surface. Our hypothesis is that the reconstructed surface from shading for PA samples might contain strong evidence of synthetic patterns in comparison to authentic samples. To the best of our knowledge, this is the first attempt at using this kind of reasoning for the PAD problem.

In contrast with 3D reconstruction and photometric stereo techniques, SfS techniques require only one image of the object under analysis. Moreover, the estimation of the shape using these techniques does not require any additional hardware, which makes possible the application of our technique in devices equipped with only an RGB camera such as smartphones and webcams. Fig. I illustrates a face surface reconstruction using an SfS algorithm [7], which will be described in details in Section III. In summary, the main contributions of this paper are:

- a new method for face presentation attack detection based on intrinsic properties of the surfaces reconstructed through shape-from-shading modeling, which allows the use of the proposed method in systems equipped with a single RGB sensor;
- a new shallow CNN network designed to learn discriminant features from the albedo, reflectance, and depth maps for the PAD problem, which achieved competitive results for intra- and inter-dataset evaluation protocols;
- the investigation of using a shape-from-shading technique for the presentation attack detection problem.

We organize the remainder of this paper as follows. Section II presents some relevant related approaches to the face presentation attack detection. Section III describes the proposed method. Section IV presents the datasets and evaluation protocols used in this paper, besides experimental results and a comparison with methods available in the literature. Finally, Section V presents the conclusions and possible directions for future work.

II. Related Work

Texture analysis is undoubtedly an important and promising line of investigation that made possible progress in this research field toward the development of effective PAD algorithms. Back to the First Competition on Counter Measures to 2D Facial Spoofing Attacks [8], the best proposed algorithms [9], [10] explored different texture descriptors, such as Local Binary Patterns (LBP), Gray-Level Co-Occurrence Matrices (GLCM), Histogram of Oriented Gradients (HOG), among others, for detecting printed-based attempted attacks [11].

In order to push the state-of-the-art further, the Second Competition on Counter Measures to 2D Face Spoofing Attacks [12] presented to the community a novel dataset (Replay-Attack dataset) [13] containing three different attack types, print-, photo-, and video-based attacks. The winner teams addressed the problem through a feature-level fusion of texture- and motion-based features. The Replay-Attack dataset was fairly challenging at the time, inviting further interesting investigations of other cues for detecting face presentation attacks.

Erdogmus and Marcel [14]–[16] explored depth information for detecting face presentation attacks by analyzing both color and depth data obtained by Microsoft's Kinect sensor. The authors proposed to use the Local Binary Patterns (LBP) descriptor in both color and depth images to produce feature vectors, which were used to feed a Linear Discriminant Analysis (LDA) classifier to reveal an attempted attack. Pinto *et al.* [4], [17], [18] also exploited alternatives for detecting face presentation attacks exploiting the residual noise present in the fake biometric sample left during their recapture and reconstruction such as blurring effects, printout artifacts, Moiré patterns, among others. Similarly, Garcia and Queiroz [19] and Wen *et al.* [20] explored these and other artifacts related to image distortions caused mainly by the recapture process of the original biometric signal.

Another cue that has been an object of investigation in the literature is regarding the reflectance of the objects. Although skin reflectance presents great variation due to different tonalities of human skin [21], [22], researchers have successfully used it in several applications [23]–[25]. In these cases, however, the reflectance is measured through extra-devices, for instance, thermal infrared imagery and near-infrared imagery. Alternatively, some computational methods for estimating the reflectance map of a scene from RGB images [26]–[28] have been proposed in the literature to decompose an RGB image into their reflectance and illumination components [29].

CNN-based techniques also have been considered in the literature. Menotti *et al.* [30] proposed a framework for optimizing CNN architectures for the PAD problem considering different modalities, including face biometrics. The authors also proposed a shallow CNN network, the SpoofNet network, for detecting iris, fingerprint, and face presentation attacks. Although the authors achieved good results using this technique, this work did not consider more challenging protocols such as inter-dataset protocols and cross-sensor setups. Other attempts at using shallow networks for the PAD problem have been reported in the literature. CLDnet [31] and ContactlensNet [32] networks were designed to detect contact lenses-based presentation attacks in iris biometric systems and contain five and two convolutional layers, respectively. In [33], the authors proposed an ensemble approach using shallow networks fed with transformed inputs, which presented good generalization in cross-domain evaluations [33]. Finally, Atoum *et al.* [34] combine a patch- and depth-based CNN for face PAD, in which the authors also achieved good results for the intra-dataset evaluation protocol. However, this work also reported their results only in the intra-dataset protocol.

Several works in the literature exploited a fine-tuning of existing deep architectures such as AlexNet, VGG, VGG-Face, and GoogleLeNet [35]–[38]. However, in general, these architectures achieved near-perfect classification results for the intra-dataset and, at same time, very poor results (close to random) for the inter-dataset protocol. Recently, Rehman *et al.* [39] proposed a new CNN-based anti-spoofing technique using the VGG-11 architecture, in which the authors reported impressive results for the intra- and inter-dataset scenario. However, a serious methodological failure described by the authors in Sec. 4.2.2 of the original paper [39], made any comparison unfeasible. As mentioned by the authors, part of the testing dataset was used to estimate the threshold τ, which was used for computing the APCER, BCPER, and HTER values. More precisely, considering the inter-dataset protocol, in which we have a training dataset and a testing dataset, the authors used the training partition contained in the test dataset for estimating the threshold τ, which obviously biased the reported results. In contrast to Rehman *et al.*, this paper and other ones published in the literature use the testing dataset only to report the performance results.

Differently from previous work in the literature, in this paper, we propose a PAD technique that takes advantage of depth, albedo, and reflectance information from RGB-images, without the necessity of any extra-device such as Microsoft's Kinect, infrared sensor or light-field devices [40], [41]. Instead of using different methods for computing each one of these components, we propose to use a shape-from-shading algorithm, which enables us to estimate these three representations from a single RGB image. Additionally, we also propose a new CNN architecture able to work in the intra- and inter-dataset scenario. To the best of our knowledge, our work is the first one to deal with these three schemes simultaneously using shape-from-shading modeling for detecting face presentation attacks.

III. Proposed Method

In this section, we present our proposed method for face PAD, which is based on intrinsic properties of the surface such as reflectance, albedo, and shape. As previously described, we propose the use of SfS for measuring these properties and use them as input for a Convolutional Neural Network (CNN) method, which learns discriminative features for detecting presentation attacks. The advantage of using an SfS method, instead of using an extra-device sensor, is two-fold: (i) a shape-from-shading method gives us an estimation of these three properties at once, at no extra cost; and (ii) we came up with a completely data-driven method, which enables our method for use in biometric systems equipped with only an RGB camera such as smart-phones.

The human ability to perceiving the shape of the objects from its shading it is one of the most important aspects of the human visual system. This ability is essential for the human understanding of the world under a three-dimensional perspective [42]. Some studies show that human can accurately use shading cues to infer changes in the surface orientation [42]–[44]. In computer vision, there are two main classes of methods for estimating the shape from shading: photometric stereo and shape-from-shading methods. An essential difference between them is that photometric stereo methods require two or more images of the same object under different lighting conditions, whereas shape-from-shading methods require only one image of the object to estimate its normal surfaces, making SfS methods very attractive to our problem [45].

We believe that some SfS methods are more appropriate to be applied in our problems than other methods in the literature, according to assumptions and restrictions imposed during the formulation of the problem. For instance, methods that add a smoothness constraint to the surface might be inadequate to be used in our problem because such constraint is not contemplated when recovering the shape of faces due to some cavities. Our work is based on Tsai's approach [7], which does not impose any restriction that could render its use improper for the PAD problem.

A. Why Shape-from-shading for Detecting Presentation Attacks?

According to the law of reflections [46], the physical mechanism of the light reflection can be characterized in terms of absorption and irradiation of the light incident onto a surface. Basically, the beam of light that affects a flat surface may be absorbed, transmitted, and reflected. The light reflected can be mathematically understood by Snell's law, which predicts the directions of the light reflected and refracted, taking into account the refraction index of the material and the roughness of its surface, that is, the smoothness or texture of the surface.

When a beam of light affects a truly flat surface, each incident ray is reflected at the same angle that we have between the surface normal and such incident ray, but on the opposite side of the surface normal. In contrast, when a beam of light affects rough surfaces, the incident light is reflected in several different directions. An ideal diffuse reflecting surface that reflects the incident light in all directions is said to exhibit a Lambertian reflection. These two processes are known as specular and diffuse reflection, respectively. Although many materials can exhibit both types of reflection, some materials reflect the light more diffusely (e.g., paper fibers, non-absorbing powder such as plaster, poly-crystalline material such as white marble, among others) [47]–[50].

The reflective power of the material is another interesting physical property that we believe to be useful for the presentation attack detection problem. This property is also known as surface albedo and can be defined as a measure of how much light incident on a surface is reflected without being absorbed. In other words, this property measures the reflectivity of a material and gives an estimate of the level of the diffuse reflection [51], [52]. Thus, objects that appear white reflect most of the incident light, indicating a high albedo, whereas dark objects absorb most of the incident light, indicating a low albedo.

Finally, the last physical property investigated in this work is the depth information associated with an object in the scene. Considering the presentation attack instruments known in the literature (e.g., photograph, video replay, mask), we clearly have a significant loss of depth information, except for mask-based presentation attacks. In fact, several works published in the literature have successfully investigated features able to characterize the depth information of face regions to point out an attempted attack [15], [16], [53]. Basically, these approaches propose to use depth sensors, such as Microsoft's Kinect sensor, to find an accurate depth map of the scene.

B. Surface Reconstruction: Recovering the Depth, Reflectance and Albedo maps

In this section, we revisited the mathematical formulation of the shape-from-shading method presented by Tsai *et al.* since such equations were directly implemented in the algorithm used in this work. Also, the following equations are the essence of the shape-from-shading formulation adopted in this work, and consequently, it is import to recap these equations to have a clear understanding of what the albedo, reflectance, and depth maps are.

The Tsai's algorithm [7] uses a linear approximation of reflectance function R to estimate the depth function Z from a single image. The main idea is to apply a discrete approximation for the surface normal using the finite differences method in order to linearize the reflectance function R in terms of Z, and then solve the linear system through the Jacobi iterative method [54].

Suppose that a point at position (x, y, z), in camera coordinates, is at a distance z from the image plane and there is a mapping between points in camera coordinates onto the image plane created using the parallel projection (not taking into account any sort of distortion). Assuming that depth information is a function of image plane coordinates $Z = Z_{x,y}$, then the change in depth δz of the point related to the change in image plane coordinate (x, y) can be expressed by using the Taylor series expansion of the function Z about point (x, y) as:

$$\delta z \approx \frac{\partial z}{\partial x}\delta x + \frac{\partial z}{\partial y}\delta y \tag{1}$$

The gradient of the surface at point (x, y, z) is the vector $(p, q) = (\frac{\partial z}{\partial x}, \frac{\partial z}{\partial y})$ and, therefore, the normal of a surface patch is related to the gradient by $\mathbf{n} = (p, q, 1)$, since the gradient vector is orthogonal to the level surface $Z_{x,y}$.

Now, suppose a Lambertian surface, which has only diffuse reflectance and the brightness is proportional to the energy of the incident light. In this case, the amount of light energy incident on a surface is proportional to the area of the surface as seen from the light source position, which can be expressed as:

$$\begin{aligned} E_{x,y} &= R(p, q) = \rho I(\mathbf{n} \cdot \mathbf{s}) \\ &\Rightarrow R(p, q) = \rho \frac{(-p, -q, 1)}{\sqrt{1 + p^2 + q^2}} \cdot \frac{(-p_s, -q_s, 1)}{\sqrt{1 + p_s^2 + q_s^2}} \end{aligned} \tag{2}$$

where $E_{x,y}$ is the intensity at pixel (x, y), I is the illuminance (or strength of light), $\mathbf{n} = (-p, -q, 1)$ is the surface normal, $\mathbf{s} = (-p_s, -q_s, 1)$ is the light source direction, and ρ is the albedo of the surface.

The SfS method employed in this work uses a discrete approximation for p and q as shown in Equation 3 and performs a linear approximation of Equation 4 based on the Taylor series expansion considering the first order terms of the function f about a given depth map Z^{n-1}, which give us a linear system of equations (Equation 5).

$$\begin{aligned} p &= \frac{\partial z}{\partial x} = Z_{x,y} - Z_{x-1,y} \\ q &= \frac{\partial z}{\partial y} = Z_{x,y} - Z_{x,y-1} \end{aligned} \tag{3}$$

$$\begin{aligned} 0 &= f(E_{x,y}, R(\partial z/\partial x, \partial z/\partial y)) \\ 0 &= E_{x,y} - R(Z_{x,y} - Z_{x-1,y}, Z_{x,y} - Z_{x,y-1}) \end{aligned} \tag{4}$$

$$\begin{aligned} 0 &= f(Z_{x,y}) \\ &\approx f(Z_{x,y}^{n-1}) + (Z_{x,y} - Z_{x,y}^{n-1}) \frac{d}{dZ_{x,y}} f(Z_{x,y}^{n-1}) \end{aligned} \tag{5}$$

When we consider $Z_{x,y} = Z_{x,y}^n$, that is, the depth at *n*-th iteration, Equation 5 can be rewritten (Equation 6) and solved by the Jacobi iterative method [54], considering an initial estimate of the depth map $Z_{x,y}^0 = 0$.

$$Z_{x,y}^n = Z_{x,y}^{n-1} + \frac{-f(Z_{x,y}^{n-1})}{\frac{df(Z_{x,y}^{n-1})}{dZ_{x,y}}} \tag{6}$$

The reflectance and albedo maps also can be obtained directly from Equation 2. After finding the depth map $Z_{x,y}^n$ at point (x, y), the reflectance map can be computed by using Equation 7, while the albedo map can be found through Equation 8.

$$R(p, q) = \max\left(0, \rho \frac{pp_s + qq_s + 1}{\sqrt{1 + p^2 + q^2}}\right) \tag{7}$$

$$\rho_{x,y}^{(n)} = \frac{I_{x,y}}{\mathbf{n}_{x,y}^{(n)} \cdot \mathbf{s}} \tag{8}$$

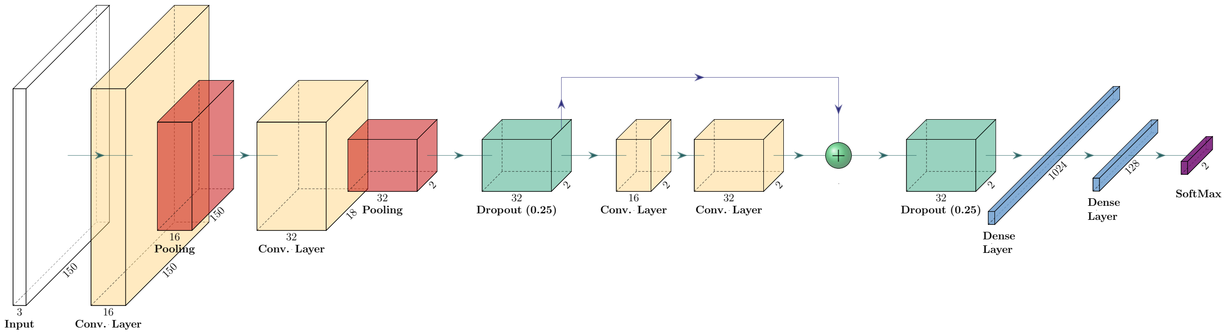


Fig. 2. Overview of the proposed method for face presentation attack detection. Given a training set, we reconstruct the face surfaces by using an SfS method, which produces estimates for the albedo, reflectance, and depth maps. Then, the proposed CNN network is trained to learn discriminative features from these maps. Finally, the classification model found during the training phase is used to decide if a given testing sample is a bona fide presentation or a presentation.

C. Convolutional Neural Network for Learning Intrinsic Surface Properties

Convolutional Neural Networks (CNNs) [55] is a well-known machine learning technique designed to learn discriminative features from input data and also a mapping function, for instances, for classification purposes. Their ability to learn an efficient and effective representation space from data has been extensively reported by the scientific community, producing impressive results in many applications such as object recognition [56], [57], video analysis [58], fake images detection [59], [60], presentation attack detection [2], [30], among others.

Inspired by Menotti *et al.* [30] and He *et al.* [61] approach, the CNN architecture proposed in this work is composed of a new variant of the SpoofNet network followed by one residual block, as illustrated in Fig. 2. The original SpoofNet is a shallow CNN architecture composed of two convolutional layers, containing 16 and 64 filters, respectively, with a kernel of size 5×5. Each convolutional layer is followed by a max-pooling layer, with a kernel of size $3{\times}3$ and a stride of 2 pixels, and by a local normalization layer with a kernel of size 9×9. On the other hand, the proposed CNN, hereafter named as SfSNet network, comprises two convolutional layers with 16 and 32 filters, respectively, with a kernel of size $3{\times}3$ and stride of 1 pixel. Each convolutional layer is followed by a max-pooling layer, with a kernel of size $9{\times}9$ and a stride of 8 pixels. Furthermore, in contrast to SpoofNet which uses RGB images as input, the proposed CNN uses reflectance, albedo, and depth maps as input to extract meaningful information for detecting presentation attacks. At the end, we use dropout regularization, with a dropout factor of 25%, to avoid overfitting. Finally, we investigated two strategies to train the proposed CNN with SfS maps:

1) using the SfS maps individually to train one CNN model for each map; and
2) using the SfS maps to compound a multi-channel input tensor to train a single CNN and thus have only one decision model. In this strategy, we computed the three SfS maps (albedo, reflectance, and depth) for each color channel available in the RGB color space and concatenated them to come up with an input tensor of $150 \times 150 \times 9$.

IV. Experimental Results

In this section, we present experimental results for the proposed method. Section IV-A describes the datasets used in the experiments, whereas Section IV-B describes the experimental protocols used to validate our approach. Section IV-C shows the experimental setup of the proposed method regarding its parameters, and Sections IV-D and IV-E show the obtained results using the maps obtained with the shape-from-shading algorithm and feature learning process. The remaining sections describe performance results considering the intra- and inter-dataset evaluation protocols and a comparison among the proposed method and other approaches reported in the literature.

A. Datasets

We evaluated the proposed method in three datasets freely available in the literature, which are described in details in the following sections:

1) Replay-Attack dataset: This dataset contains videos of presentation attacks and bona fide presentations of 50 identities, which were recorded with a webcam with a pixel resolution of 320×240. This dataset provides three types of presentation attacks: print-, mobile- and video- attacks with high-definition resolution, which were split into three subsets: the training set with 360 videos; the development set containing 360 videos; and testing set with 480 videos, totaling $1,000$ videos of presentation attacks and 200 videos of bona fide presentation [13].

2) CASIA dataset: This dataset comprises 600 videos of presentation attacks and bona fide accesses of 50 identities. The authors recorded both presentation attack and bona fide presentation videos in three different qualities: (i) low-quality videos captured by an old USB camera with 480×640 pixel resolution; (ii) normal-quality videos, which were recorded by a new USB camera with 480×640 pixel resolution; and (iii) high-quality videos captured with a Sony NEX-5 camera with $1,920 \times 1,080$ pixel resolution. The types of presentation attacks contained in this dataset include warped photo attacks, cut photo attacks, photos and video attacks. Finally, this dataset provides 240 videos for training and 360 videos for testing, totaling 150 videos of bona fide presentations and 450 videos of presentation attacks [62].

3) UVAD dataset: This dataset contains bona fide presentation and presentation attack videos of 404 identities, all created at Full HD quality. The videos were recorded in two sections considering different illumination conditions and environments. In total, this dataset provides $16,268$ presentation attack videos and 808 bona fide presentation videos, which were recorded through six acquisition sensors of different manufacturers (Sony, Kodak, Olympus, Nikon, Canon, and Panasonic). The video attacks were simulated with seven different display devices, also with HD and Full HD quality. The authors recommend using the videos from Sony, Kodak and Olympus sensors for training, and the videos from Nikon, Canon and Panasonic sensors for testing. This evaluation protocol provides $3,872$ videos for training and $6,416$ videos for testing, totaling 404 bona fide presentation videos and $9,884$ presentation attack videos [18].

B. *Experimental Protocols*

The performance of the proposed method is assessed through two metrics recommended by ISO/IEC 30107-3 [63], Attack Presentation Classification Error Rate (APCER) and the Bona fide Presentation Classification Error Rate (BPCER), in which the APCER is the proportion of presentation attacks incorrectly classified as bona fide presentations and the BPCER is the proportion of bona fide presentations incorrectly classified as presentation attacks. Although, the ISO/IEC does not aggregate these two measures, in this work we additionally use two measures for that, the Equal Error Rate (EER) and Half Total Error Rate (HTER), since the evaluation protocol for some datasets recommends using them. The EER value is defined by the threshold for which the APCER and BPCER rates are equal, and the HTER is the average of APCER and BPCER measures computed in a threshold τ, which must be defined in a development set.

We evaluated our approach under two experimental protocols, the intra- and inter-dataset protocols. In the intra-dataset scenario, we validate the proposed method using each dataset separately, and we follow the official protocols defined for each dataset in their original papers. Therefore, the Replay-Attack dataset is comprised of three subsets: the training set, which was used to fit a classification model; the development set used to find the EER threshold; and the test set, which was used only to report the APCER, BPCER, and HTER values. For the datasets composed of two subsets (CASIA and UVAD), we use the training set to fit a classification model and to find the EER threshold, and the test set to report the final results in terms of APCER, BPCER, and HTER. We also reported the EER value obtained in the test set for the CASIA dataset, as suggested by the dataset's authors. In the inter-dataset setup, we use one dataset for training the proposed method and a different one to test it.

C. *Experimental Setup*

This section describes parameter configurations and implementation details of the proposed method for reproducibility purposes of the results presented in this paper.

Regarding the shape-from-shading algorithm used in this work, the only parameter required by this algorithm is the light source direction, whose value has been set to coordinate $(0, 0, 1)$. Therefore, we considered that the primary light source is perpendicular to the faces during the acquisition, which is a reasonable choice taking into account the datasets used in the experiments and the nature of the PAD problem. As the shape-from-shading algorithm works upon images, we subsample the videos to have about 61 frames per video ($\approx$ 2 seconds). Moreover, we apply the shape-from-shading algorithm to each color channel (RGB representation) with the frames cropped in the face regions, whose locations were provided by the datasets' authors. Finally, we resize the SfS maps found by the algorithm to (150×150), which were used to feed the CNN networks.

We conducted the training process of the CNN networks using 150 epochs and batches of 64. We used the Adadelta solver for minimizing the categorical cross-entropy objective function using a learning rate of $1\mathrm{e}{-2}$ without the learning decay strategy. Finally, we use an L2 regularization in the soft-max classifier, whose value was configured to $1\mathrm{e}{-4}$. The seeds were pre-defined in order to obtain reproducibility of our results. Finally, the class decision (bona fide presentation vs. presentation attack) for an input video was taken considering the fusion scores of its 61 frames by computing the median. We use Keras (version 2.1.3) and Tensorflow (version 1.4.1) frameworks[1] to implement the proposed CNN network and the source code of all proposed methods are freely available[2].

D. *Evaluation of the Proposed CNN Architecture*

In this section, we evaluate the CNN network proposed in this work, which was inspired by the SpoofNet [30], a shallow network designed for the PAD problem, and by the Residual Networks (ResNet). Here, we show the effectiveness of our proposed CNN, the SfSNet network, by comparing it with the original SpoofNet, ResNet [61], and Xception networks [64]. For both Xception and ResNet networks, we performed a fine-tuning of a pre-trained version trained upon the ImageNet dataset [55] since we do not have enough data for training them from scratch for the PAD problem. Thus, after loading the pre-trained weights, we remove the top layer and we freeze the remaining layers to indicate that such layers will not be trained. Thereafter, we add a fully connected layer with $1,024$ units followed by a soft-max layer with 2 outputs.

Table I shows a comparison among these CNN networks for the CASIA dataset using the intra-dataset protocol. Both SpoofNet and SfSNet networks outperform ResNet and Xception networks. We believe that shallow networks are more suitable for the PAD problem due to the nature of patterns to be learned by the networks, which came from artifacts added to the synthetic samples such as blurring, banding effect, Moiré patterns, among others. Noticeably, such patterns can be better understood as low-level features and deeper networks are suitable for learning high-level features such as part of complex objects.

[1] https://keras.io and https://www.tensorflow.org

[2] The source code is freely available for scientific purposes on GitHub (https://github.com/allansp84/shape-from-shading-for-face-pad).

TABLE I
PERFORMANCE RESULTS (IN %) FOR THE CASIA DATASET CONSIDERING THE INTRA-DATASET EVALUATION PROTOCOL.

Architecture	Map Type	APCER	BPCER	HTER	Mean HTER
	Albedo	68.9	68.9	68.9	
ResNet [61]	Depth	34.8	48.9	41.9	58.2
	Reflectance	65.6	62.2	63.9	
	Albedo	8.5	78.9	43.7	
Xception [64]	Depth	18.2	55.6	36.9	38.1
	Reflectance	29.6	37.8	33.7	
	Albedo	8.2	14.4	11.3	
SpoofNet [30]	Depth	14.4	11.1	12.8	11.1
	Reflectance	8.5	10.0	**9.3**	
	Albedo	6.7	8.9	**7.8**	
SfSNet	**Depth**	11.1	5.6	**8.3**	**8.6**
(Proposed Method)	Reflectance	15.2	4.4	9.8	

E. CNNs with Shape-from-Shading Maps as Inputs

TABLE II
PERFORMANCE RESULTS (IN %) OF SFSNET NETWORK FOR REPLAY-ATTACK AND CASIA DATASETS CONSIDERING THE INTRA-DATASET EVALUATION PROTOCOL.

Dataset	Map Type	APCER	BPCER	HTER
	Albedo	11.0	5.0	8.0
	Depth	5.3	0.0	2.6
Replay-Attack	Reflectance	6.5	1.3	3.9
	Majority Vote	0.0	6.8	3.4
	Multi-channel input tensor	6.3	0.0	**3.1**
	Albedo	6.7	8.9	7.8
	Depth	11.1	5.6	8.3
CASIA	Reflectance	15.2	4.4	9.8
	Majority Vote	3.3	7.4	5.4
	Multi-channel input tensor	1.5	3.3	**2.4**

In this section, we evaluate two strategies to extract meaningful information from the different maps using the proposed CNN network. The experiments presented in this section were performed using the intra-dataset evaluation protocol. The first strategy consists of training a CNN network for each one of the three types of maps available (albedo, reflectance and depth maps), which give us three CNN-based classifiers. Then, a fusion approach based on the majority vote is employed in order to have a final score to decide whether a testing sample is a presentation attack or a genuine access. The second approach consists of giving to the network the multi-channel input tensor as described in Section III. Table II shows the obtained results considering these two strategies.

According to the results, the multi-channel input tensor outperform the majority vote fusion strategy with a relative error reduction of 7.7% for the Replay-Attack dataset and more than 50.0% for the CASIA dataset. Besides having a significant reduction overall, once we need to train only one model, the multi-channel input tensor strategy also facilitate the training of our CNN-based classifier and is more efficient. This is because different maps may behave as a data augmentation approach towards avoiding possible problems regarding over-fitting. We also notice a significant difference in performance of some maps across different datasets. For instance, the albedo presented comparable performance, while the depth and reflectance estimations presented considerable differences. We believe that physical properties of devices used to build the datasets can potentially introduce these variations in performances for some maps such as reflectance and albedo (e.g., matte and glossy monitors). Regarding the depth map, we notice that light direction estimation used to compute the normal surface can introduce a significant error in depth estimation.

F. Intra-dataset Evaluation Protocol

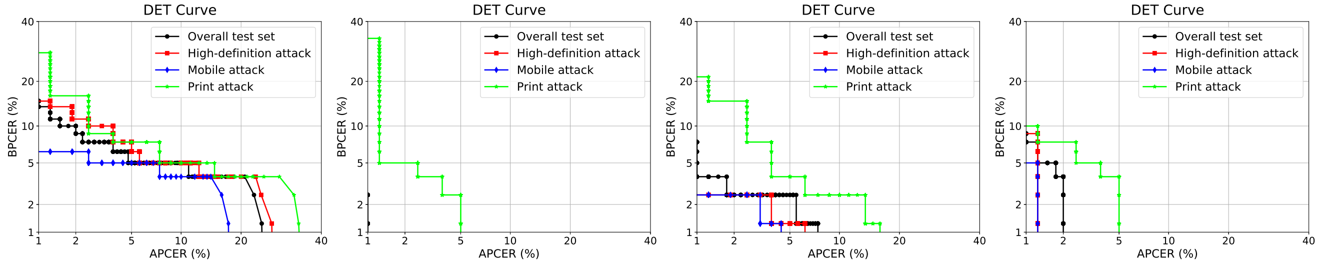
In this section, we present performance results of our approach for the datasets considered in this work. We followed the evaluation protocol defined for each dataset and we also reported performance results using the metrics suggested by the datasets' authors.

1) Replay-Attack Dataset: Fig. 3 shows the obtained Detection Error Tradeoff (DET) curves for different maps obtained by the shape-from-shading algorithm and for the three types of presentation attacks contained in this dataset. The aim of this experiment is investigating the discriminability of these maps for detecting the different attack types. The results indicate that mobile-based presentation attacks were the most easily detected by the proposed algorithm. Considering the depth map (Fig. 3(b)), the proposed approach achieved an HTER of 2.6% considering the overall test set and perfect BPCER rates for all attack types. Table III shows the performance results for the network trained using the depth maps.

TABLE III
PERFORMANCE RESULTS (IN %) FOR THE REPLAY-ATTACK DATASET CONSIDERING THE PRESENTATION ATTACKS SIMULATIONS INDIVIDUALLY.

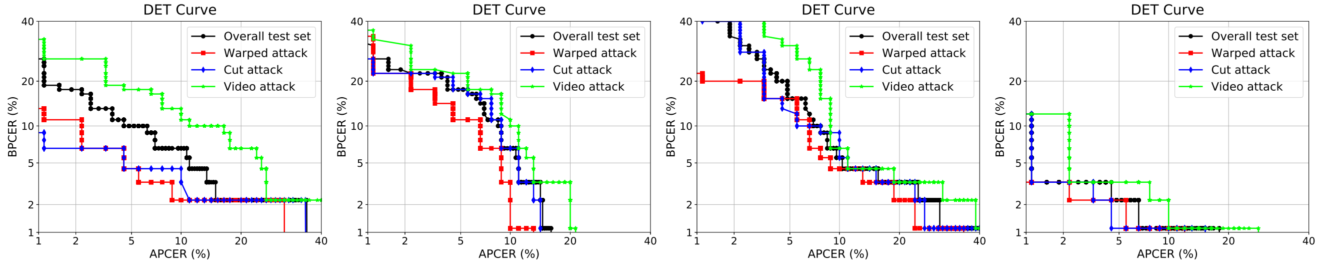
Attack Type	APCER	BPCER	HTER
Hight-Def	4.4	0.0	2.2
Mobile	**3.1**	**0.0**	**1.6**
Print	11.3	0.0	5.6
Overall test	5.3	0.0	2.6

2) CASIA Dataset: Fig. 4 illustrates obtained DET curves considering the different presentation attack simulations. Here,



(a) Albedo maps (b) Depth maps (c) Reflectance maps (d) Multi-channel input tensor

Fig. 3. Results obtained on Replay-Attack dataset for the three attack types and for the different maps obtained with the shape-from-shading algorithm.



(a) Albedo maps (b) Depth maps (c) Reflectance maps (d) Multi-channel input tensor

Fig. 4. Results obtained on CASIA dataset for the different attack types and for the different maps obtained with the shape-from-shading algorithm.

the network trained with the multi-channel input tensor achieved the best performance results for all categories of attack present in this dataset. Furthermore, the warped- and cut-based attacks were easier to detect than video-based attempted attacks. We also notice that the network trained with the multi-channel input tensor showed more robustness to deal with the different types of presentation attack. Table IV shows the error rates for this network, which achieved an HTER of 2.4%. For the warped attack simulations, we achieved an APCER rate of 0.0%, which means the network detected all warped photo attack simulations.

TABLE IV
PERFORMANCE RESULTS (IN %) FOR THE CASIA DATASET CONSIDERING THE PRESENTATION ATTACKS SIMULATIONS INDIVIDUALLY.

Attack Type	APCER	BPCER	HTER	EER
Warped photo	**0.0**	**3.3**	**1.7**	**2.2**
Cut photo	1.1	3.3	2.2	3.3
Video	3.3	3.3	3.3	3.3
Overall test	1.5	3.3	2.4	3.3

3) UVAD Dataset: In this section, we evaluate the proposed method in a challenging scenario with presentation attacks and bona fine presentations, both captured with different sensors, which is named in the literature as a cross-sensor scenario [65]. Table V shows the obtained results for the different maps, which shows that network trained using the depth maps is the most discriminative network for detecting the presentation attack in this dataset. Although the HTER of 14.5% obtained in this dataset is higher than the ones in previous datasets, this result is the lowest achieved in the literature as shown in Section IV-H.

TABLE V
PERFORMANCE RESULTS (IN %) FOR THE UVAD DATASET CONSIDERING THE DIFFERENT MAPS OBTAINED WITH THE SHAPE-FROM-SHADING ALGORITHM.

Map Type	APCER	BPCER	HTER
Albedo	24.6	20.0	22.3
Depth	**10.7**	**18.3**	**14.5**
Reflectance	22.1	31.7	26.9
Multi-channel input tensor	12.4	21.7	17.0

G. Inter-dataset Evaluation Protocol

We now turn our attention to the obtained results for the inter-dataset evaluation protocol, which is the most challenging evaluation protocol nowadays. The difficulty of this evaluation protocol raises up from the fact that we have training and testing scenarios different in terms of acquisition sensors, light conditions, and environment (e.g., different background).

Table VI shows the obtained results of the proposed method trained with the CASIA dataset and tested upon the other ones previously mentioned, beside the OuluNPU dataset [2], [66] considering its hardest evaluation protocol (Protocol IV). Surprisingly, the proposed method achieved an outstanding performance result for the Replay-attack dataset when we consider multi-channel input tensor and only the video-based attempted attack videos for training our CNN network, with an HTER of 9.8%. For both OuluNPU and UVAD datasets, our method achieved a better performance when we consider depth maps for training our CNN network. On the other side, our method achieved an APCER, BPCER, and HTER of 34.8%, 24.4%, and 29.6%, respectively, by using the Replay-Attack dataset for training and the CASIA dataset for testing and considering reflectance maps. Finally, considering the UVAD

TABLE VI
RESULTS (IN %) OBTAINED WITH THE CROSS-DATASET PROTOCOL CONSIDERING BOTH PRESENTATION ATTACKS SIMULATIONS INDIVIDUALLY AND THE OVERALL TEST SETS OF REPLAY, UVAD, AND OULUNPU DATASETS .

Training Set	**Testing Sets**								
CASIA	**Replay-Attack**			**OuluNPU**			**UVAD**		
	APCER	**BPCER**	**HTER**	**APCER**	**BPCER**	**HTER**	**APCER**	**BPCER**	**HTER**
Video	10.8	8.8	**9.8**	67.9	4.2	36.0	57.9	21.7	39.8
Overall	8.3	51.3	29.8	49.6	12.5	31.0	34.8	36.7	35.7
Warped	65.8	27.5	46.6	41.7	13.3	**27.5**	36.8	30.0	**33.4**
Cut	92.0	2.5	47.3	75.4	8.3	41.9	58.1	23.3	40.7

TABLE VII
COMPARISON AMONG EXISTING CNN-BASED METHODS CONSIDERING THE INTRA- AND INTER-BASED EVALUATION PROTOCOLS FOR THE DATASETS CONSIDERED IN THIS WORK.

Methods	**Intra-Dataset Protocol**			**Inter-Dataset Protocol**	
	Replay-Attack	**CASIA**		**Replay-Attack**	**CASIA**
	HTER	**EER**	**HTER**	**HTER**	**HTER**
Li et al. [35] (Fine-tuned VGG-Face)	4.3	5.2	–	–	–
Li et al. [35] (DPCNN)	6.1	4.5	–	–	–
Atoum et al. [34] (Patches and Depth-Based CNNs)	0.7	2.7	2.3	–	–
Menotti et al. [30] (Architecture Optimization)	0.8	–	–	–	–
Li et al. [67] (Hybrid CNNs)	1.6	2.2	–	–	–
Pinto et al. [36] (Fine-tuned VGG network)	0.0	–	6.7	49.7	47.2
Yang et al. [37] (Fine-tuned AlexNet)	2.7	–	6.3	41.4	42.0
Patel et al. [38] (GoogLeNet + Eye-Blink Detection)	0.5	–	–	12.4	31.6
Wang et al. [68] (Adversarial Domain Adaptation)	1.4	3.2	–	6.6	37.8
Liu et al. [69]	–	–	–	27.6	28.4
SfSNet (Proposed Method)	3.1	3.3	2.4	9.8	29.6

dataset for training and the CASIA dataset for testing, the proposed method achieved an APCER, BPCER, and HTER values of 66.7%, 12.2%, and 39.4%, respectively, using the depth maps. We believe the variabilities, in terms of attack types and video quality, present in the CASIA dataset were essential for the proposed method to achieved better results upon the Replay-Attack dataset, which also contain different attack types such as printed- and video-based attempted attacks. On the other hand, the UVAD dataset contemplate only video-based attempted attacks.

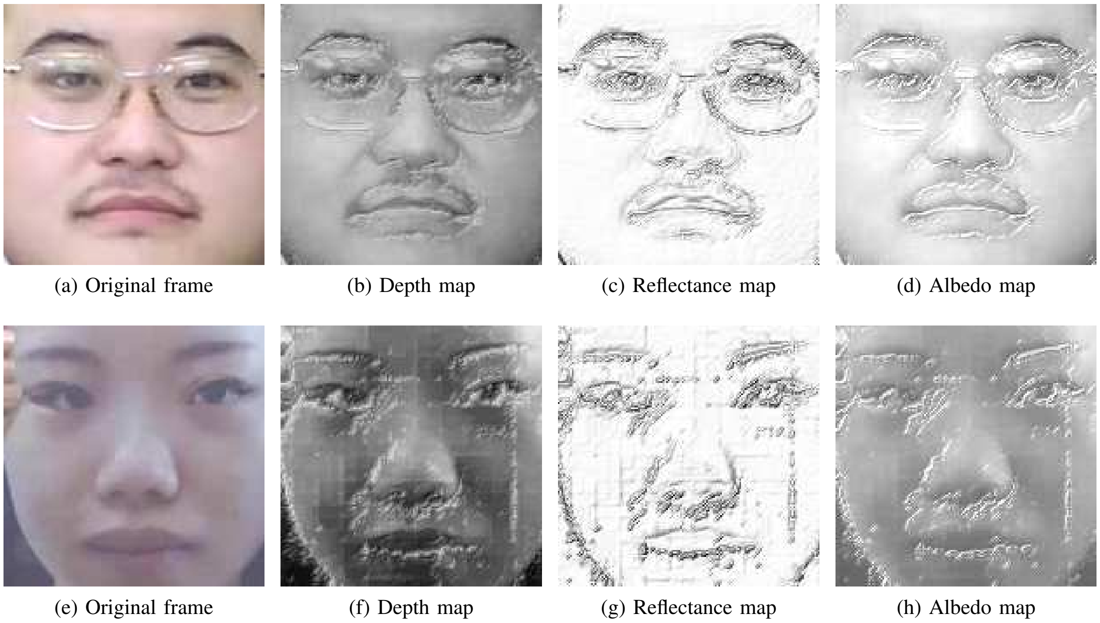
H. Comparison with State-of-the-Art Methods

In this section, we compare the proposed method with other methods available in the literature. We select the most effective CNN networks designed for the PAD problem, including the networks specifically designed to estimate depth maps from RGB images without using any kind of extra device [34]. We notice that most effective methods that take into account the intra-dataset evaluation protocol achieved poor performance results in the inter-dataset protocol, as shown in Table VII. The proposed method achieved the lowest HTER for the inter-dataset protocol and competitive results for the intra-dataset evaluation protocol, which demonstrates the potential of the proposed method. Considering the complexity of the existing networks, i.e. the number of convolutional layers, the proposed CNN architecture provides a reasonable trade-off between performance and hardware requirement, which can be directly translated into memory consumption and training time of the network.

I. Visual Assessment

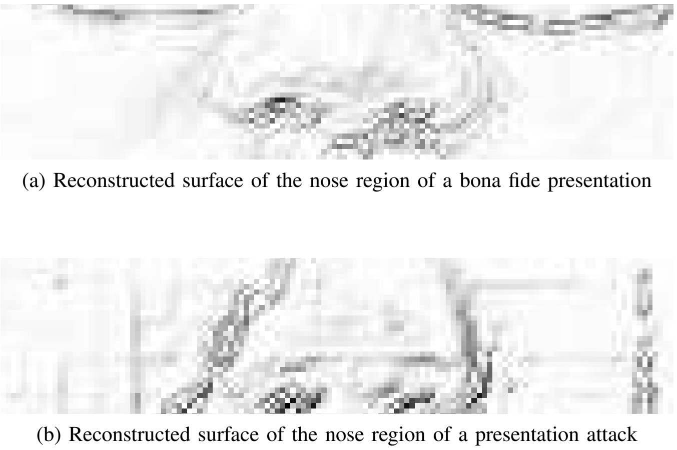
In this section, we show a visual assessment of the albedo, reflectance, and depth maps. Fig. 5 depicts these maps computed from a bona fide presentation and from a presentation attack video frame. These examples illustrate how the artifacts affect the reconstruction of the surface, specially, in this example, of the depth and reflectance maps. We believe that the way how the algorithm computes the depth might improve the highlighting of the artifacts present in the presentation attack images. As mentioned in Section III-B, we perform the estimation of the depth locally, which means that each point (x, y) is reconstructed interdependently. When the algorithm tries to compute the first and second order derivative of outliers (e.g., noise or printing artifact), we come up with a situation where the approximation might not be applied, which produces the white spots in the reconstructed maps. Moreover, the first and second order derivative computations can potentially highlight printing artifacts, i.e., horizontal and vertical lines. Fig. 6 shows the details of the reconstructed surface considering a video frame of both classes of the PAD problem, in which we evince natural pattern for the genuine access (e.g., skin roughness) and synthetic patterns for presentation attack image (e.g., horizontal and vertical lines).

Also, the reliance on a shallow network, which requires less data for estimating good values for its parameters is also an important factor that help us to find reliable models with higher generalization capabilities. The problem of training deep CNNs using small training datasets can potentially trigger overfitting problems and thus produce unreliable models. To


(a) Original frame (b) Depth map (c) Reflectance map (d) Albedo map

(e) Original frame (f) Depth map (g) Reflectance map (h) Albedo map

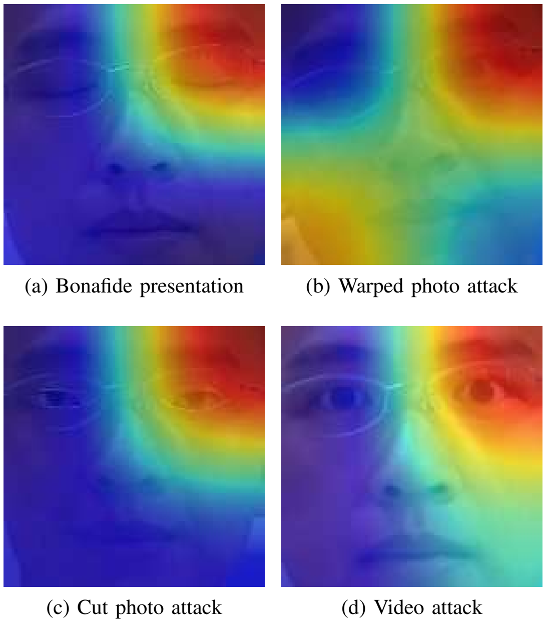
Fig. 5. Example of a bona fide presentation video frame (first row) and presentation attack video frame (second row). The first column illustrates original frames captured by the acquisition sensor, whereas the other columns show their respective maps, in which the some artifacts unseen in the original frame (horizontal and vertical lines) were highlighted during reconstruction.


(a) Reconstructed surface of the nose region of a bona fide presentation

(b) Reconstructed surface of the nose region of a presentation attack

Fig. 6. Details of the reconstructed surface for the video frames showed in Fig. 5 from a genuine access (a) and an attempted attack (b), in which we found strong evidence of a natural (skin roughness) texture pattern and of a synthetic (horizontal and vertical lines) texture pattern for these respective classes.

overcome this problem, this research takes advantage of transformed feature spaces and leverages shallow networks to learn useful representations for PAD detection on such transformed inputs, thus requiring less training data examples. Finally, we used the Gradient-weighted Class Activation Mapping (Grad-CAM) algorithm [70] to produce a visual explanation to our model, as illustrated in Figure 7. In this experiment, we first performed a prediction using a reflectance map as input. Next, we computed the Grad-CAM, which produced a Heatmap that highlights the importance of an image's regions for the final decision-making. Finally, we combined the Heatmap and the original image to visualize the importance of raw image regions. We could observe that our CNN examines different regions of an input image to detect the different types of presentation attack.


(a) Bonafide presentation (b) Warped photo attack

(c) Cut photo attack (d) Video attack

Fig. 7. Visual explanation obtained using the Gradient-weighted Class Activation Mapping (Grad-CAM) algorithm, considering our CNN network trained with reflectance maps. Figure 7 (a) shows the visual explanation to a bonafide presentation, while the Figures 7-(b-d) illustrated the visual explanation for different attack types.

V. Conclusion and Future Work

In this paper, we proposed an algorithm for detecting presentation attacks based on intrinsic properties of the scene such as albedo, reflectance, and depth of the scene. We showed that these properties are useful for detecting different types of presentation attacks with satisfactory results in terms of error rates. We also proposed a novel CNN network specially designed for learning features from these different maps. The ability of CNN networks in learning from data was crucial for our method to achieve the reported results, since the handcrafting feature engineering of these different maps could be much more challenging.

The experimental results corroborated the effectiveness of our CNN networks trained using these different maps. Particularly, the network trained with the depth maps and with the concatenated maps showed more robustness for detecting presentation attacks considering the inter-dataset evaluation protocol. For the intra-dataset evaluation protocol, the depth map achieved the best performance results for the UVAD and Replay-Attack datasets, whereas the concatenated maps achieved the best performance results for the CASIA dataset. We believe there could be some complementarity between these maps, which would allow our CNN network to learn good features and deal with this complex dataset that contains several kinds of photo and video presentation attacks.

Unquestionably, the inter-dataset evaluation protocol was the hardest scenario for the proposed method, even considering the cross-sensor scenario, in which we achieved better results than the state-of-the-art, as confirmed through the results obtained for the UVAD dataset. We believe our work could help the community to have a better understanding about this challenging problem, since the proposed method was able to spot strong evidences of presentation attacks considering the photographs- and video-based attempted attacks in the reconstructed surface of the faces.

Future research efforts include the investigation of alternative approaches to combining the albedo, reflectance, and depth maps toward extracting complementary patterns. This is useful for detecting presentation attacks, as well as the investigation of new approaches to recovering the surface properties from shading by taking into account other reflectance models such as Bidirectional reflectance distribution function (BRDF). The study of methods for finding the light source that operates in a real scenario (not with synthetic images) could also be a promising investigation path toward improving the facial surface reconstruction.

Acknowledgment

We thank CAPES (DeepEyes project), FAPESP (grant #2019/16253-1), FAPESP through the grants #2017/12631-6 and #2017/12646-3 (DéjàVu project), CNPq (Grant #423797/2016-6) and Microsoft Research for the financial support.

References

[1] S. Marcel, M. S. Nixon, and S. Z. Li, *Handbook of Biometric Anti-Spoofing: Trusted Biometrics Under Spoofing Attacks*. Springer Publishing Company, Incorporated, 2014.

[2] Z. Boulkenafet, J. Komulainen, Z. Akhtar, A. Hadid *et al.*, "A competition on generalized software-based face presentation attack detection in mobile scenarios," in *IEEE Int. Joint Conf. on Biometrics*, Oct. 2017, pp. 688–696.

[3] A. Liu, J. Wan, S. Escalera, H. Jair Escalante, Z. Tan, Q. Yuan, K. Wang, C. Lin, G. Guo, I. Guyon, and S. Z. Li, "Multi-modal face anti-spoofing attack detection challenge at cvpr2019," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[4] A. Pinto, H. Pedrini, W. R. Schwartz, and A. Rocha, "Face spoofing detection through visual codebooks of spectral temporal cubes," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 4726–4740, Dec. 2015.

[5] Z. Boulkenafet, J. Komulainen, and A. Hadid, "Face anti-spoofing based on color texture analysis," in *IEEE Int. Conf. Image Process.*, Sep. 2015, pp. 2636–2640.

[6] B. K. Horn, "Shape from shading: A method for obtaining the shape of a smooth opaque object from one view," Massachusetts Institute of Technology, Cambridge, MA, USA, Tech. Rep., 1970.

[7] T. Ping-Sing and M. Shah, "Shape from shading using linear approximation," *Image Vis. Comput.*, vol. 12, no. 8, pp. 487–498, 1994.

[8] M. Chakka, A. Anjos, S. Marcel, Tronci *et al.*, "Competition on counter measures to 2-d facial spoofing attacks," in *IEEE Int. Joint Conf. on Biometrics*, 2011, pp. 1–6.

[9] J. Määttä, A. Hadid, and M. Pietikäinen, "Face spoofing detection from single images using micro-texture analysis," in *IEEE Int. Joint Conf. on Biometrics*, Oct. 2011, pp. 1–7.

[10] W. Robson Schwartz, A. Rocha, and H. Pedrini, "Face spoofing detection through partial least squares and low-level descriptors," in *IEEE Int. Joint Conf. on Biometrics*, Oct. 2011, pp. 1–8.

[11] A. Anjos and S. Marcel, "Counter-measures to photo attacks in face recognition: A public database and a baseline," in *IEEE Int. Joint Conf. on Biometrics*, Oct. 2011, pp. 1–7.

[12] I. Chingovska, J. Yang, Z. Lei, D. Yi *et al.*, "The 2nd Competition on Counter Measures to 2D Face Spoofing Attacks," in *IAPR Int. Conf. Biometrics*, Jun. 2013, pp. 1–6.

[13] I. Chingovska, A. Anjos, and S. Marcel, "On the effectiveness of local binary patterns in face anti-spoofing," in *Int. Conf. Biometrics Special Interest Group*, Sep. 2012, pp. 1–7.

[14] N. Erdogmus and S. Marcel, "Spoofing 2D face recognition systems with 3D masks," in *Int. Conf. Biometrics Special Interest Group*, 2013, pp. 1–8.

[15] ——, "Spoofing in 2D face recognition with 3D masks and anti-spoofing with kinect," in *IEEE Int. Conf. Biometrics: Theory Appl. and Syst.*, Sep. 2013, pp. 1–6.

[16] ——, "Spoofing face recognition with 3d masks," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 7, pp. 1084–1097, Jul. 2014.

[17] A. Pinto, H. Pedrini, W. R. Schwartz, and A. Rocha, "Video-based face spoofing detection through visual rhythm analysis," in *Conf. Graph., Patterns and Images*, Aug. 2012, pp. 221–228.

[18] A. Pinto, W. Robson Schwartz, H. Pedrini, and A. Rocha, "Using visual rhythms for detecting video-based facial spoof attacks," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 5, pp. 1025–1038, May 2015.

[19] D. C. Garcia and R. L. de Queiroz, "Face-Spoofing 2D-Detection Based on Moiré-Pattern Analysis," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 4, pp. 778–786, Apr. 2015.

[20] D. Wen, H. Han, and A. K. Jain, "Face spoof detection with image distortion analysis," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 4, pp. 746–761, Apr. 2015.

[21] S. R. Marschner, S. H. Westin, E. P. F. Lafortune, K. E. Torrance, and D. P. Greenberg, "Reflectance measurements of human skin," Program of Computer Graphics, Cornell University, Tech. Rep. PCG-99-2, 1999.

[22] C. C. Cooksey, D. W. Allen, and B. K. Tsai, "Reference data set of human skin reflectance," *J. Res. Natl. Inst. Stand. Technol.*, vol. 122, pp. 1–5, 2017.

[23] J. Han and B. Bhanu, "Human activity recognition in thermal infrared imagery," in *IEEE Int. Conf. Comput. Vision and Pattern Recognit.*, Jun. 2005, pp. 17–17.

[24] C. Dai, Y. Zheng, and X. Li, "Pedestrian detection and tracking in infrared imagery using shape and appearance," *Comput. Vis. Image Underst.*, vol. 106, no. 2, pp. 288–299, 2007.

[25] S. Z. Li, R. Chu, S. Liao, and L. Zhang, "Illumination invariant face recognition using near-infrared images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 4, pp. 627–639, Apr. 2007.

[26] B. K. P. Horn and R. W. Sjoberg, "Calculating the reflectance map," *Appl. Opt*, vol. 18, no. 11, pp. 1770–1779, Jun. 1979.

[27] R. L. Cook and K. E. Torrance, "A reflectance model for computer graphics," *ACM Trans. Graph.*, vol. 1, no. 1, pp. 7–24, Jan. 1982.

[28] N. Almoussa, "Variational retinex and shadow removal," University of California, Department of Mathematics, Tech. Rep., 2009.

[29] N. Kose and J.-L. Dugelay, "Reflectance analysis based countermeasure technique to detect face mask attacks," in *Int. Conf. Digital Signal Process.*, 2013, pp. 1–6.

[30] D. Menotti, G. Chiachia, A. Pinto, W. R. Schwartz, H. Pedrini, A. X. Falcão, and A. Rocha, "Deep representations for iris, face, and fingerprint spoofing detection," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 4, pp. 864–879, Apr. 2015.

[31] P. Silva, E. Luz, R. Baeta, H. Pedrini, A. X. Falcao, and D. Menotti, "An approach to iris contact lens detection based on deep image

representations," in *2015 28th SIBGRAPI Conference on Graphics, Patterns and Images*. IEEE, 2015, pp. 157–164.

[32] R. Raghavendra, K. B. Raja, and C. Busch, "Contlensnet: Robust iris contact lens detection using deep convolutional neural networks," in *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 2017, pp. 1160–1167.

[33] A. Kuehlkamp, A. Pinto, A. Rocha, K. W. Bowyer, and A. Czajka, "Ensemble of multi-view learning classifiers for cross-domain iris presentation attack detection," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 6, pp. 1419–1431, June 2019.

[34] Y. Atoum, Y. Liu, A. Jourabloo, and X. Liu, "Face anti-spoofing using patch and depth-based cnns," in *IEEE Int. Joint Conf. on Biometrics*, Oct. 2017, pp. 319–328.

[35] L. Li, X. Feng, Z. Boulkenafet, Z. Xia, M. Li, and A. Hadid, "An original face anti-spoofing approach using partial convolutional neural network," in *Int. Conf. on Image Process. Theory, Tools and Appl*, Dec. 2016, pp. 1–6.

[36] A. Pinto, H. Pedrini, M. Krumdick, B. Becker, A. Czajka, K. W. Bowyer, and A. Rocha, *Deep Learning in Biometrics*. CRC Press, 2018, ch. Counteracting Presentation Attacks in Face Fingerprint and Iris Recognition, p. 49.

[37] J. Yang, Z. Lei, and S. Z. Li, "Learn convolutional neural network for face anti-spoofing," *CoRR*, vol. abs/1408.5601, 2014.

[38] K. Patel, H. Han, and A. K. Jain, "Cross-database face antispoofing with robust feature representation," in *Biometric Recognition*, Z. You, J. Zhou, Y. Wang, Z. Sun, S. Shan, W. Zheng, J. Feng, and Q. Zhao, Eds. Cham: Springer International Publishing, 2016, pp. 611–619.

[39] Y. A. U. Rehman, L. M. Po, and M. Liu, "Livenet: Improving features generalization for face liveness detection using convolution neural networks," *Expert Systems with Applications*, vol. 108, pp. 159–169, 2018.

[40] C. Galdi, V. Chiesa, C. Busch, P. Correia, J. Dugelay, and C. Guillemot, "Light fields for face analysis," *Sensors*, vol. 19, no. 12, pp. 2687–2687, June 2019.

[41] A. George, Z. Mostaani, D. Geissenbuhler, O. Nikisins, A. Anjos, and S. Marcel, "Biometric face presentation attack detection with multi-channel convolutional neural network," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 42–55, 2020.

[42] V. S. Ramachandran, "Perception of shape from shading," *Nature*, vol. 331, pp. 163–166, Jan. 1988.

[43] A. P. Pentland, "Finding the illuminant direction," *J. Opt. Soc. Am.*, vol. 72, no. 4, pp. 448–455, Apr. 1982.

[44] P. Winston and B. Horn, *The psychology of computer vision*, ser. McGraw-Hill computer science series. McGraw-Hill, 1975.

[45] R. Szeliski, *Computer Vision: Algorithms and Applications*, 1st ed. New York, NY, USA: Springer-Verlag New York, Inc., 2010.

[46] P. Tipler and G. Mosca, *Physics for Scientists and Engineers*, ser. Physics for Scientists and Engineers. W. H. Freeman, 2007.

[47] K. E. Torrance and E. M. Sparrow, "Theory for off-specular reflection from roughened surfaces," *J. Opt. Soc. Am.*, vol. 57, no. 9, pp. 1105–1114, Sep. 1967.

[48] M. Kazayawoko, J. J. Balatinecz, and R. T. Woodhams, "Diffuse reflectance fourier transform infrared spectra of wood fibers treated with maleated polypropylenes," *J. Appl. Polym. Sci.*, vol. 66, no. 6, pp. 1163–1173, 1997.

[49] M. P. Fuller and P. R. Griffiths, "Diffuse reflectance measurements by infrared fourier transform spectrometry," *Anal. Chem.*, vol. 50, no. 13, pp. 1906–1910, 1978.

[50] K. Moradi, C. Depecker, and J. Corset, "Diffuse reflectance infrared spectroscopy: Experimental study of nonabsorbing materials and comparison with theories," *Appl. Spectrosc.*, vol. 48, no. 12, pp. 1491–1497, Dec. 1994.

[55] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of the 25th*

[51] T. Hyodo, *Radiation physics: proceedings of the International Symposium on Radiation Physics*, ser. NBS special publication. U.S. Dept. of Commerce, National Bureau of Standards, 1977, ch. Backscattering of Gamma Rays, pp. 110–118.

[52] Q. Zheng and R. Chellappa, "Estimation of illuminant direction, albedo, and shape from shading," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, no. 7, pp. 680–702, Jul. 1991.

[53] G. Goswami, M. Vatsa, and R. Singh, "Rgb-d face recognition with texture and attribute features," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 10, pp. 1629–1640, Oct. 2014.

[54] Y. Saad, *Iterative Methods for Sparse Linear Systems*, 2nd ed. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2003. *Int. Conf. Neural Inf. Process. Syst.*, ser. NIPS'12, vol. 1. USA: Curran Associates Inc., 2012, pp. 1097–1105.

[56] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *IEEE Int. Conf. Comput. Vision and Pattern Recognit.*, Jun. 2015, pp. 1–9.

[57] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 386–397, Feb 2020.

[58] M. Perez, S. Avila, D. Moreira, D. Moraes, V. Testoni, E. Valle, S. Goldenstein, and A. Rocha, "Video pornography detection through deep learning techniques and motion information," *Neurocomputing*, vol. 230, pp. 279–293, 2017.

[59] E. R. de Rezende, G. C. Ruppert, A. Theóphilo, E. K. Tokuda, and T. Carvalho, "Exposing computer generated images by using deep convolutional neural networks," *Signal Process.: Image Commun.*, 2018.

[60] T. Pomari, G. Ruppert, E. Rezende, A. Rocha, and T. Carvalho, "Image splicing detection through illumination inconsistencies and deep learning," in *IEEE Int. Conf. Image Process.*, 2018 - To appear.

[61] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Int. Conf. Comput. Vision and Pattern Recognit.*, Jun. 2016, pp. 770–778.

[62] Z. Zhang, J. Yan, S. Liu, Z. Lei, D. Yi, and S. Li, "A face antispoofing database with diverse attacks," in *IAPR Int. Conf. Biometrics*, Apr. 2012, pp. 26–31.

[63] ISO/IEC 30107-3:2017, *Information technology – Biometric presentation attack detection – Part 3: Testing and reporting*, 2017.

[64] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *IEEE Int. Conf. Comput. Vision and Pattern Recognit.*, Jul. 2017.

[65] D. Yambay, B. Becker, N. Kohli *et al.*, "Livdet iris 2017 - iris liveness detection competition 2017," in *IEEE Int. Joint Conf. on Biometrics*, Oct 2017, pp. 733–741.

[66] Z. Boulkenafet, J. Komulainen, L. Li, X. Feng, and A. Hadid, "Oulu-npu: A mobile face presentation attack database with real-world variations," in *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, May 2017, pp. 612–618.

[67] L. Li, Z. Xia, L. Li, X. Jiang, X. Feng, and F. Roli, "Face anti-spoofing via hybrid convolutional neural network," in *Int. Conf. Frontiers and Advances in Data Sci.*, Oct. 2017, pp. 120–124.

[68] G. Wang, H. Han, S. Shan, and X. Chen, "Improving cross-database face presentation attack detection via adversarial domain adaptation," in *2019 International Conference on Biometrics (ICB)*, June 2019, pp. 1–8.

[69] Y. Liu, A. Jourabloo, and X. Liu, "Learning deep models for face anti-spoofing: Binary or auxiliary supervision," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2018, pp. 389–398.

[70] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017, pp. 618–626.