Iris Liveness Detection Using a Cascade of Dedicated Deep Learning Networks

Juan E. Tapia, *Member, IEEE*, Sebastian Gonzalez, *Member, IEEE*, and Christoph Busch, *Member, IEEE*

***Abstract*— Iris pattern recognition has significantly improved the biometric authentication field due to its high stability and uniqueness. Such physical characteristics have played an essential role in security applications and other related areas. However, presentation attacks, also known as spoofing techniques, can bypass biometric authentication systems using artefacts such as printed images, artificial eyes, textured contact lenses, etc. Many liveness detection methods that improve the robustness of these systems have been proposed. The first International Iris Liveness Detection competition, where the effectiveness of liveness detection methods is evaluated, was first launched in 2013, and its latest iteration was held in 2020. In this paper, we present the approach that won the LivDet-Iris 2020 competition using two-class scenarios (bona fide iris images vs. presentation attack iris images). Additionally, we propose new three-class and four-class scenarios that complement the competition results. These methods use a serial architecture based on a MobileNetV2 modification, trained from scratch to classify bona fide iris images versus presentation attack images. The bona fide class consists of live iris images, whereas the attack presentation instrument classes consist of cadaver, printed, and contact lenses images, for a total of four species. All the images were pre-processed and weighted per class to present a fair evaluation. This approach is primarily focused on detecting the bona fide class over improving the detection of presentation attack instruments. For the two, three, and four classes scenarios $BPCER_{10}$ values of 0.99%, 0.16%, and 0.83% were obtained respectively, whereas for the $BPCER_{20}$ values of 3.09%, 0.16%, and 3.77% were obtained, with the best model overall being the proposed 3-class serial model. This work reaches competitive results according to the reported results in the LivDet-Iris 2020 competition.**

***Index Terms*— LiveDet, PAD, presentation attack detection, MobileNet.**

I. Introduction

IRIS recognition systems has been shown to be robust over time, affordable, non-invasive, and touchless; these strengths will allow it to grow in the market in the coming years [1]. Iris recognition systems are usually based on near-infrared (NIR) lighting and sensors, and have been shown to be susceptible to Presentation Attack Instruments (PAI) [2], where PAI refers to a biometric characteristic or object used in a presentation attack. Presentation Attack Detection (PAD) refers to the ability of a biometric system to recognize PAIs, that would otherwise fool the system into recognizing an illegitimate user as a genuine one, by means of presenting a synthetic forged version of the original biometric trait to the capture device. The biometric community, including researchers and vendors, have thrown themselves into the challenging task of proposing and developing efficient protection mechanisms against this threat [3], where PAD methods have been suggested as a solution to this vulnerability. Attacks are not restricted to merely theoretical or academic scenarios anymore, as they are starting to be carried out against real-life operations. One example is the hacking of Samsung Galaxy S8 devices with the iris unlock system, using a regular printer and a contact lens. This case has been reported to the public from hacking groups attempting to get recognition for real criminal cases, including from live biometric demonstrations at conferences.[1] An ideal PAD technique should be able to detect all of these attacks, along with any new or unknown PAI species that may be developed in the future [4]. PAD for iris recognition systems is a very dynamic topic, as it has been shown in past editions of the LivDet competition, revealing that there are still open problems to get efficient methods for usage in capturing devices. This paper contributes to improving the state of the art, adds a new database and also explains the methodology used for the winning team.

In order to improve PAD methods, a few competitions and databases have been created, such as the LivDet-Iris.[2] The goal of the Liveness Detection Competition (LivDet-Iris) is to compare biometric liveness detection methodologies, using a standardized testing protocol and large quantities of attack presentation (spoofed) and bona fide presentation samples. This competition has shown that there are still challenges for the detection of iris presentation attacks, mainly when unknown materials or capture devices are used to generate the attacks [5]. The results show that even with latest advances in presentation attacks, printed iris PAIs, as well as

Manuscript received May 27, 2021; revised August 17, 2021 and October 12, 2021; accepted November 16, 2021. Date of publication December 3, 2021; date of current version December 16, 2021. This work was funded by the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE and TOC Biometrics Company. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. P. C. Yuen. *(Corresponding author: Juan E. Tapia.)*

Juan E. Tapia and Christoph Busch are with the da/sec-Biometrics and Internet Security Research Group, Hochschule Darmstadt, 64295 Darmstadt, Germany (e-mail: juan.tapia-farias@h-da.de; christoph.busch@h-da.de).

Sebastian Gonzalez is with the Research and Development Center SR-226, TOC Biometrics Company, Santiago 7510099, Chile (e-mail: sebastian.gonzalez@tocbiometrics.com).

Digital Object Identifier 10.1109/TIFS.2021.3132582

[1] https://www.forbes.com/sites/ianmorris/2017/05/23/samsung-galaxy-s8-iris-scanner-hacked-in-three-simple-steps/?sh=2ad04efaccba

[2] https://livdet.org/

patterned contact lenses PAIs, are still difficult for software-based systems to detect according with the quality of the images. In LivDet-2017 [5], printed iris images were easier to be differentiated from bona fide images in comparison to patterned contact lenses, as it was also shown in the previous competitions. Some properties of the samples (images) are unknown during training, making the challenge a difficult task, as the winning algorithm did not recognize from 11% to 38% of the attack images, depending on the database. Therefore, the PAD techniques are still an open challenge in NIR, and it has been even less explored in VIS periocular images and multiple capture devices.

The results from the LivDet-2020 [6] competition indicate that iris PAD is still far from a fully solved research problem. Large differences in accuracy among baseline algorithms, which were trained with significantly different data, stress the importance of access to large and diversified training datasets, encompassing a large number of PAI species. The winning team (our method) also achieved the lowest Bona Fide Classification Error Rate (BPCER) of 0.46%, out of all nine algorithms in the three categories. This aligns well with the operational goal of PAD algorithms to correctly detect bona fide presentations (i.e., and not to contribute to system's False Non-Match Rate), and capture as many attacks as possible.

One of the main challenges to improve PAD systems is the quantity and quality of the data available. Printed images are easy to reproduce with different kinds of paper. Conversely, post-mortem images [7], and PAI species such as contact lenses, cosmetic lenses, plastic lenses, all sourced from different brands, are hard to get. Therefore a subject-disjoint dataset containing different iris patterns is difficult to achieve. Alternatively, these datasets can be synthetically created using deep learning techniques [8], [9].

In this work, a serial, two-stage architecture for classification of bona fide, presentation attack, high-quality printed, and digitally displayed images of LivDet-2020, plus three complementary databases were explored using deep learning techniques. The main contributions of this work can be summarized a follows:

- *Architecture*: A serial, two-stage architecture is proposed. This consists of a modified MobileNetV2 model ("MobileNetv2a"), trained from scratch, which is utilized to differentiate between bona fide presentation and presentation attack. A second MobileNet named "MobileNetv2b", trained from scratch for four scenarios, which is then used to detect printed/contact-lenses/cadaver impostor attacks by identifying the physical source of the images. See Figure 3.
- *Network inputs*: A strong set of experiments of serial and parallel structures of DNNs was evaluated with two, three, and four classes, using NIR images. Bona fide versus contact lenses, print-out, cadavers, electronics and prosthetic displays were used as input to the network. Also, separate and exhaustive experiments were realized using one of these four types of input, and the results were analyzed.
- *Weights*: Balanced class weights were used in order to correctly represent the number of images per class. Most of the spoofing databases are unbalanced according to PA scenarios. Weighted classes help to balance the dataset and to get realistic results.
- *Database*: This paper presents two new databases, one database to increase the number of bona fide images (10,000), and a second database to increase the number of printed PAIs with high-quality images (1,800). Both databases will be available to other researchers upon request, for research purposes only (See Section III).
- *Data-Augmentation (DA)*: An aggressive DA technique to train the modified MobileNetV2a and MobileNetV2b networks was used. These images allow the network to learn more challenging scenarios considering blurring, Gaussian noise, coarse occlusion, crop, and others.
- *Winning method*: Focused on the correct classification of bona fide images instead of the identification of several PAI species, the serial approach presented in this work reached first place in the LivDet-2020 competition. Furthermore, the current proposal with three and fourth classes complement our results presented in the competition, featuring more challenging scenarios.
- *Not self-reported*: The two-stage algorithm presented in this paper was evaluated by the organizers of the competition in an independent test on unknown data; the test data was not available for the participants.

II. Related Work

Multiple PAD methods for iris recognition systems have been proposed in the scientific literature, given the increased adoption of these systems for a variety of different operations, which increases the threats of attacks on these sensitive systems.

Zou *et al.* [10] have presented a novel algorithm, 4DCycle-GAN, for expanding spoofed iris image databases, by synthesizing artificial iris images wearing textured contact lenses. The proposed 4DCycle-GAN follows the Cycle Consistent Adversarial Networks (Cycle-GAN) framework, which translates between one kind of image (bona fide iris images) to another kind (textured contact lenses iris images). Despite the improvements on Conditional Generative Adversarial Networks, there are still some open problems that limit its application for image generation. Therefore, the method helps to create and increase the number of images based on conditional GANs while preserving the information in the images of each PAI in the NIR spectrum.

Hu *et al.* [11] investigated the use of regional features in iris PAD (RegionalPAD). Features are extracted from local neighborhoods, based on spatial pyramid (multi-level resolution) and relational measures (convolution on features with variable-size kernels). Several feature extractors, such as Local Binary Patterns (LBP) [12], Local Phase Quantization (LPQ) [13], and intensity correlogram are examined. They used a three-scale LBP-based feature, since it achieves the best performance, as pointed out by the original authors.

Gragnaniello *et al.* [14] proposes that the sclera region also contains important information about iris liveness (SIDPAD). Hence, the authors extract features from both the iris and sclera regions. The two regions are first segmented, and

scale-invariant local descriptors (SID) are applied. A bag-of-feature method is then used to accumulate the features. A linear Support Vector Machine (SVM) is used to perform the final prediction. Also, in [15], domain-specific knowledge of iris PAD is incorporated into the design of their model (DACNN). With the domain knowledge, a compact network architecture is obtained, and regularization terms are added to the loss function to enforce high-pass/low-pass behavior. The authors demonstrate that the method can detect both face and iris presentation attacks.

SpoofNets [16] are based on GoogleNet, and consist of four convolutional layers and one inception module. The inception module is composed by layers of convolutional filters of dimensions 1×1, 3×3, and 5×5, executed in parallel. It has the advantage of reducing the complexity and improving the efficiency of the architecture, once the filters of dimension 1×1 help reduce the number of features before executing layers of convolution with filters of higher dimensions.

Boyd *et al.* [17] chose the ResNet50 architecture as a backbone to explore whether iris-specific feature extractors perform better than models trained for non-iris tasks. They demonstrated three types of networks: off-the-shelf networks, fine-tuned, and networks trained from scratch, with five different sets of weights for iris recognition. They concluded that fine-tuning an existing network to the specific iris domain performed better than training from scratch.

Yadav *et al.* [18], used a combination of handcrafted and deep-learning-based features for iris PAD. They fused multi-level Haralick features with VGG16 features to encode the iris textural patterns. The VGG16 features were extracted from the last fully connected layer, with a size of 4,096, and then reduced to a lower dimensional vector by Principal Component Analysis (PCA).

Nguyen *et al.* [19] proposed a PAD method by combining features extracted from local and global iris regions. First, they trained multiple VGG19 [20] networks from scratch for different iris regions. Then, the features were separately extracted from the last fully connected layer, before the classification layer of the trained models. The experimental results showed that the PAD performance was improved by fusing the features based on both feature-level and score-level fusion rules.

Kuehlkamp *et al.* [21] propose an approach for combining two techniques for iris PAD: CNNs and Ensemble Learning. Extensive experimentation was conducted using the most challenging datasets publicly available. The experiments included cross-sensor and cross-dataset evaluations. Results show a varying ability for different BSIF+CNN representations to capture different aspects of the input images. This method outperform the results presented in the LivDet-Iris 2017 competition.

Our approach, presented in the LivDet-Iris 2020 competition, reached the first place with an Average Classification Error Rate (ACER) of 29.78%. This method achieved also the lowest Bona Fide Classification Error Rate (BPCER) of 0.46% out of all nine algorithms in the three categories. This paper shows the relevance of focusing mainly on the bona fide images as a "first-filter". However, a broad space for improvement was detected in the identification of the PAI species, specially in cadaver and printed iris images. An Attack Presentation Classification Error Rate (APCER) of 9.87% was reached for the electronic display PAI species, which is lower than all competing algorithms (53.08% and 83.95%) by a large margin.

Based on previous results, this current paper proposes a new framework to improve the detection performance of PAIs per species in order to get a strong PAD method.

The rest of the article is organized as follows: Section II summarizes the related works on Presentation Attack Detection. The ISO metrics are explained in Section IV. The database description is explained in Section III. The experimental framework is then presented in Section V, and the results are discussed in Section VI. We conclude the article in Section VII.

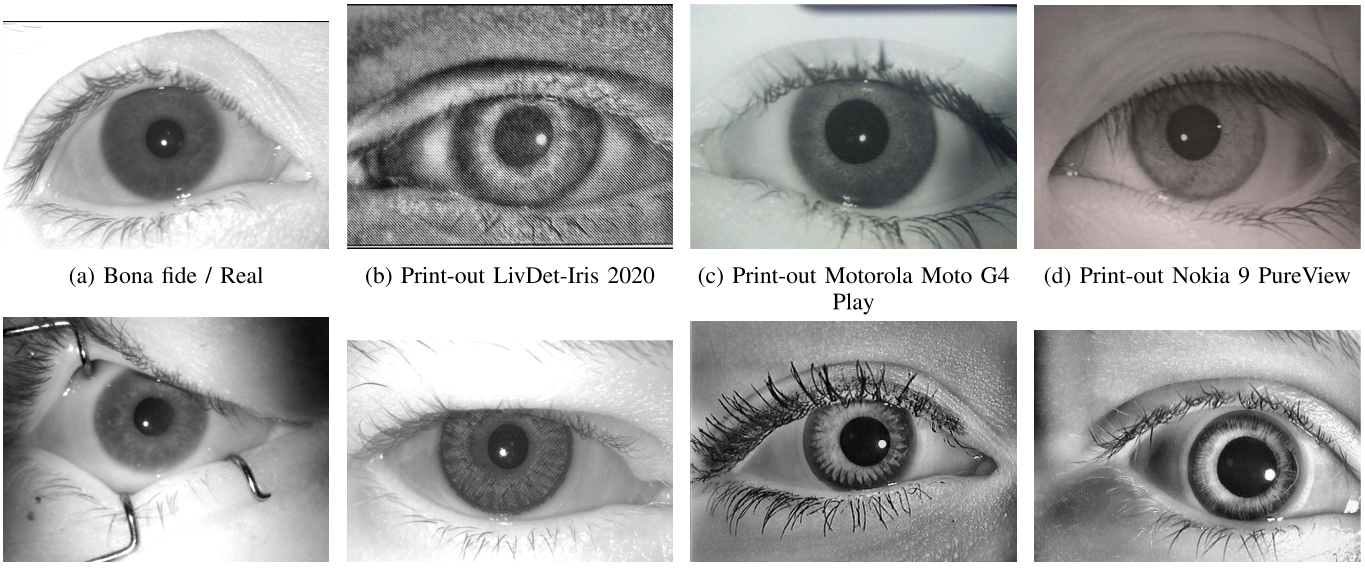
III. Databases

For this work, the LivDet-Iris 2020 competition database was used. In addition, three sets of complementary databases of iris images were also utilized. First, a database of NIR bona fide images, captured using an Iritech TD100 iris sensor with a resolution of 640×480 pixels, called "Iris-CL1". A second database, called "iris-printed-CL1", containing high-quality presentation attack images of printed PAIs was created. The goal of this database is to increase the challenge of the printed irises species, due to the noticeable visible patterns in the printed images from the LivDet-Iris 2020 database, which makes them trivial to distinguish from bona fide images. See Figure 1. The iris-printed-CL1 database contains 1,800 images captured with two smartphone devices (900 images each one): a Nokia 9 PureView device, with an image resolution of 1280×957 pixels, and a Motorola Moto G4 Play device, with an image resolution of 1280×960 pixels. Only the red channel was used. These new datasets will be available to others researchers upon request. Figure 1 present new images of printed species.

The third database is the Warsaw-BioBase-Post-Mortem-Iris v3.0 database [7]. This database contains a total of 1,094 NIR images (collected with an IriShield M2120U), and 785 visible-light images (obtained with Olympus TG-3), collected from 42 post-mortem subjects. This database was not fully available for the competition.

The LivDet-Iris 2020 database included five different PAI species, each with a different level of challenge: printed eyes, textured contact lens, electronic display, fake/prosthetic eyes, printed with add-ons, and a small number of cadaver eyes. The printed image dataset is of a very low resolution. No specific training dataset was prepared for the competition. A total of 11,918 images were made available.

The competition was different from previous editions in regards to the training dataset. the participants were encouraged to use all the data available to them (both publicly available and proprietary) to make their solutions as effective and robust as possible. The entirety of previous LivDet-Iris benchmarks were also made publicly available [5], [22], [23]. Additionally, the competition organizers shared five examples of each PAI (samples which were not used later in evaluations) to help the competitors familiarize themselves with the test



(a) Bona fide / Real (b) Print-out LivDet-Iris 2020 (c) Print-out Motorola Moto G4 Play (d) Print-out Nokia 9 PureView

(e) Cadaver (post-mortem subject) eye (f) Cosmetic contact lens (g) Electronic display (h) Patterned eye

Fig. 1. Example images of all presentation attack instruments in the database. Images c, and d show examples of the new PAI species included.

TABLE I

TRAINING DATASET SUMMARY – 11,918 IMAGES. FA, PRD, PR, REPRESENT: FAKE/PROSTHETIC DISPLAY AND PRINTED ADD-ONS. BP: BONA FIDE PRESENTATION. AP: ATTACK PRESENTATION

Dataset	Class	PAI species	Num Images
LDet-Iris-2013-Clarkson	BP	—	516
LDet-Iris-2015-Clarkson	BP	—	541
LDet-Iris-2017-Clarkson	BP	—	3,954
LDet-Iris-2020-Clarkson	BP	—	5
LDet-Iris-2020-Notre Dame	BP	—	5
LDet-Iris-2013-Clarkson	AP	Cont. Lens	840
LDet-Iris-2015-Clarkson	AP	Cont. Lens	824
LDet-Iris-2017-Clarkson	AP	Cont. Lens	1,887
LDet-Iris-2020-Notre Dame	AP	Cont. Lens	5
LDet-Iris-2015-Clarkson	AP	Printouts	1,077
LDet-Iris-2017-Clarkson	AP	Printouts	2,254
LDet-Iris-2020-Clarkson	AP	Printouts	1
LDet-Iris-2020-Clarkson	AP	Elec. Display	1
LDet-Iris-2020-Clarkson	AP	Fa, PrD, Pr	3
LDet-Iris-2020-Warsaw	AP	Cadaver Eyes	5
Total			**11,918**

TABLE II

SUMMARY OF THE NEW, COMPLETE DATABASE, WITH 27,964 IMAGES DIVIDED IN TRAIN, TEST, AND VALIDATION

Class	PAI species	Train	Val	Test	Num Im.	Sensors
BP	—	6,694	1,062	5,773	13,530	LG4000 AD 100 iCam 700 TD100
AP	Cadaver	448	531	754	1,773	IriTech IriShield
AP	Cont. Lenses Textured	3,583	900	3,244	7,727	LG4000 AD 100 iCam 700 TD100 MotoG4 Gplay
AP	Printed Prosthetic Display	4,090	1,896	2,305	8,291	Iris ID iCAM700
Total		**11,810**	**4,384**	**11,770**	**27,964**	

data format (pixel resolution, bits per pixel used to code the intensity, etc.).

Table I shows a summary of the all databases available for training in LivDet-Iris 2020. The datasets of presentation attack instruments (PAIs) were specifically created for the development of PAD methods. With the evolving of PAIs, the datasets include new challenges. A detailed technical summary of the available datasets can be found in [17], [24]. It is essential to point out that the test dataset was sequestered by the organizers and was not available for the competitors.

Table II shows a summary of all datasets available from LivDet-Iris 2020, plus Cadaver images, iris-CL1, and iris-printed-CL2. The new total count of images available is 27,964. This is more than two times the number of images shown in Table I.

A. Data Augmentation

An aggressive data augmentation (DA) method was applied when training the modified MobileNetV2 networks. All the images were normalized using a histogram equalization algorithm. A large number of images, with several operations such as affine transformations, perspective transformations, contrast changes, Gaussian noise, random dropout of image regions, cropping/padding, and blurring were included in the train dataset. These DA operations are based on the imgaug library [25], which is optimized for high performance. This

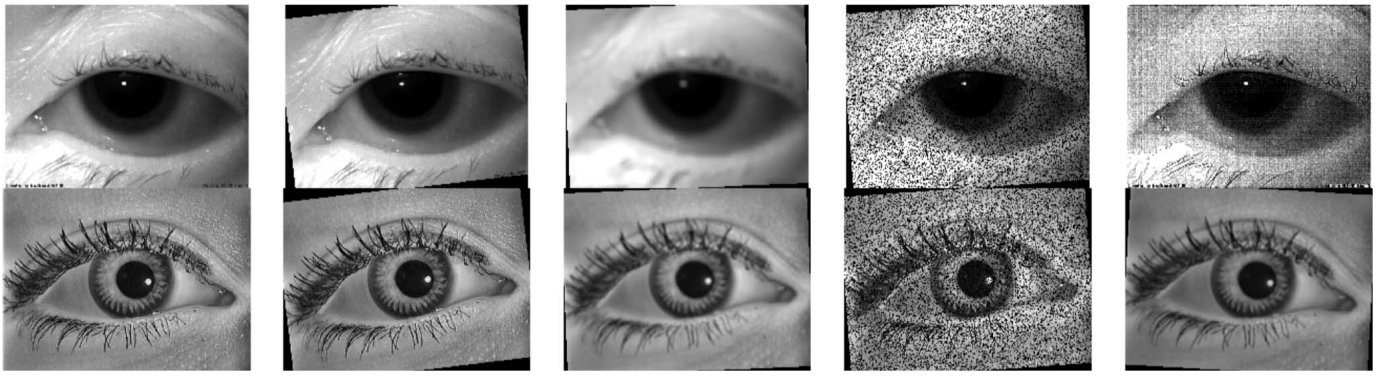


Fig. 2. Examples of the aggressive data augmentation applied randomly to bona fide and attack presentation images. Left: original images, rotation, blurring, Gaussian noise filter, and Filter Edge Enhance.

improves the quality of the training results by using very challenging images. Examples of some augmented images are shown in Figure 2.

IV. Metrics

The ISO/IEC 30107-3 standard[3] presents methodologies for the evaluation of the performance of PAD algorithms for biometric systems. The APCER metric measures the proportion of attack presentations—for each different PAI—incorrectly classified as bona fide (genuine) presentations. This metric is calculated for each PAI, where ultimately the worst-case scenario is considered. Equation 1 details how to compute the APCER metric, in which the value of N_{PAIS} corresponds to the number of attack presentation images, where RES_i for the ith image is 1 if the algorithm classifies it as an attack presentation (spoofed image), or 0 if it is classified as a bona fide presentation (real image) [26].

$$APCER_{PAIS} = 1 - (\frac{1}{N_{PAIS}}) \sum_{i=1}^{N_{PAIS}} RES_i \tag{1}$$

Additionally, the BPCER metric measures the proportion of bona fide (live images) presentations mistakenly classified as attacks presentations to the biometric capture device, or the ratio between false rejection to total genuine attempts. The BPCER metric is formulated according to equation 2, where N_{BF} corresponds to the number of bona fide (live) presentation images, and RES_i takes identical values of those of the APCER metric.

$$BPCER = \frac{\sum_{i=1}^{N_{BF}} RES_i}{N_{BF}} \tag{2}$$

These metrics effectively measure to what degree the algorithm confuses presentations of spoofed images with real images, and vice versa. Furthermore, the Average Classification Error Rate (ACER) is also used. This is computed by averaging the APCER and BPCER metrics, as shown in equation 3. Whereas the ACER metric evaluates the overall system performance, it has been deprecated in the ISO/IEC 30107-3, and is computed mainly for the purpose comparing with the state of the art. The APCER, BPCER, and ACER metrics are dependent on a decision threshold.

$$ACER = \frac{APCER + BPCER}{2} \tag{3}$$

A Detection Error Trade-off (DET) curve is also reported for all the experiments. In the DET curve, the Equal Error Rate (EER) value represents the trade-off when the APCER is equal to the BPCER. Values in this curve are presented as percentages. Additionally, two different operational points are reported, according to ISO/IEC 30107-3. BPCER_{10} which corresponds to the BPCER when the APCER is fixed at 10%, and BPCER_{20} which is the BPCER when the APCER is fixed at 5%. BPCER_{10} and BPCER_{20} are independent of decision thresholds.

V. Methodology

In this section, we introduce our baseline by utilizing a fine-tuned network, and a new network trained from scratch. Then, the detailed description of the used convolutional layers is presented.

A. Networks

MobileNetV2 [27] is based on a streamlined architecture to build lightweight deep neural networks. This allows for usage in environments with limited resources, such as mobile applications, while achieving state-of-the-art performance for tasks such as classification. MobileNetV2 trades the basic Depthwise Separable Convolution building block of MobileNetV1 [28] for a Bottle Residual block, introducing a 1×1 Expansion layer which increases the dimensionality of the input tensor before passing it to the lightweight 3×3 Depthwise Convolution layer to filter the features. Finally, the 1×1 Pointwise Convolution layer is changed for a 1×1 Projection layer, which bottlenecks the data by projecting it into a tensor of lower dimensionality. This effectively allows the basic Bottle Residual block to apply its filtering step on a high dimensional tensor given by the Expansion layer, while outputting a low dimensional tensor using the Projection layer. Furthermore, MobileNetV2 adds inverted residual connections between the bottlenecks, inverting the

[3]https://www.iso.org/standard/67381.html

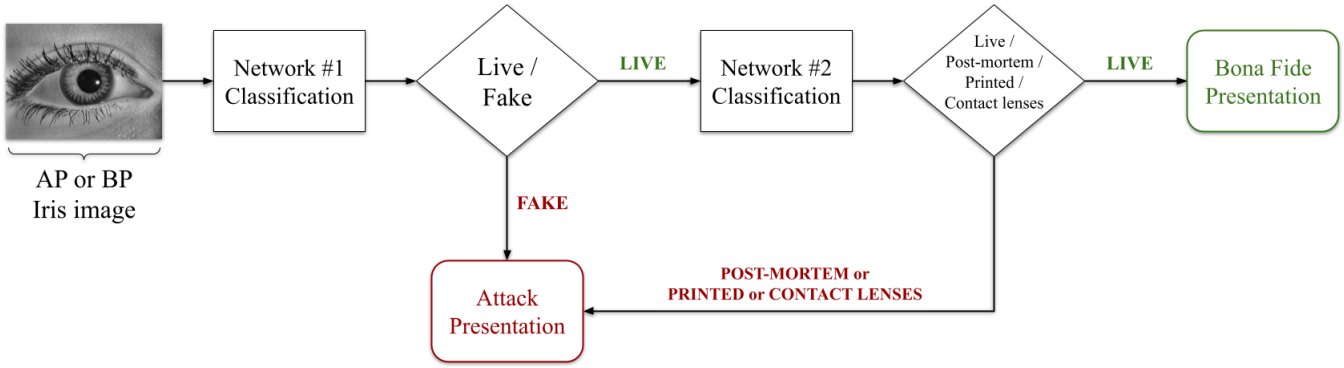


Fig. 3. Proposed Two-stage serial framework for Presentation Attack Detection. AP: Attack Presentation. BP: Bona fide Presentation. The system has one input to the serial framework.

concept proposed by networks such as ResNet [29]. These MobileNetV2 features help to speed up feature learning and improve network accuracy over its predecessor, while also reducing the amount of parameters of the network. In this work, a modified MobileNetV2 was used to detect bona fide and attack presentation images.

For this work, ImageNet [30] weights were initially used for transfer learning. However, the results of fine-tuning the network would worsen proportionally to the amount of layers that were frozen. Therefore training the networks from scratch resulted in a better classification performance overall. This is explained in more detail in Section VI.

In addition, two different network architectures are used in this work: MobileNetV2a and a modified MobileNetV2b. MobileNetV2a was trained from scratch, based on bona fide and fake species only, whereas MobileNetV2b was introduced based on bona fide, patterned contact lenses, printed, and cadaver species. See Figure 3.

B. Image Pre-Processing

All the images in the database were pre-processed using a contrast limited adaptive histogram equalization algorithm (See sub-section V-C) to improve the gray-scale intensity. Later, a weighted factor per each class was applied (See sub-section V-D). Also, a higher number of filters was applied, using the MobileNetV2 alpha parameter, from the standard 1.0 up to 1.4. Both methods are leveraged to create a two-stage classifier that can detect bona fide and attack presentation scenarios. All the images were resized to 224×224 and 448×448 according to the experiments.

C. Contrast Limited Adaptive Histogram Equalization (CLAHE)

In order to improve the quality of the images and highlight texture-related features, the CLAHE algorithm was applied. This algorithm divides an input image into an M×N grid. Afterwards, it applies equalization to each cell in the grid, enhancing global contrast and definition in the output image. All the images were divided in 8×8 sized cells.

D. Class Weights

A weight factor was estimated for each class according to the number of images of the class, helping to balance the database. Class weights are applied to the loss function, this favours under-representation and penalizes over-representation of classes by re-scaling the gradient steps during training. See Equation 4.

$$Weight_i = \frac{Nsamples}{Nclasses \times samples_i} \tag{4}$$

where $Weight_i$ is the weight for class i, $Nsamples$ is the total number of images in the database, $Nclasses$ is the total number of classes in the database, and $samples_i$ is the number of samples of class i. The weight values associated to each class are the following:

- Class 0, Cadaver: 4.4162
- Class 1, Bona fide: 0.5787
- Class 2, Pattern: 1.0133
- Class 3, Printed: 0.9443

E. Network Parameters

The number of trainable parameters and number of multiply-adds can be modified by using MobileNetV2's alpha parameter, which increases/decreases the number of filters in each layer of the network. This alpha value is known as the depth multiplier in the original MobileNet implementation:

- If $alpha = 1.0$, the default number of filters from the original MobileNet paper are used at each layer.
- If $alpha < 1.0$, proportionally decreases the number of filters in each layer.
- If $alpha > 1.0$, proportionally increases the number of filters in each layer.

For the experiments in Section VI, the value of the alpha parameter was set between 1.0 or 1.4, depending on the experiment. Furthermore, two image input sizes were tested: 224×224 and 448×448. The networks for all experiments were trained with a limit of 200 epochs. Categorical cross-entropy was used as the loss function. Adam optimization [31] was also utilized.

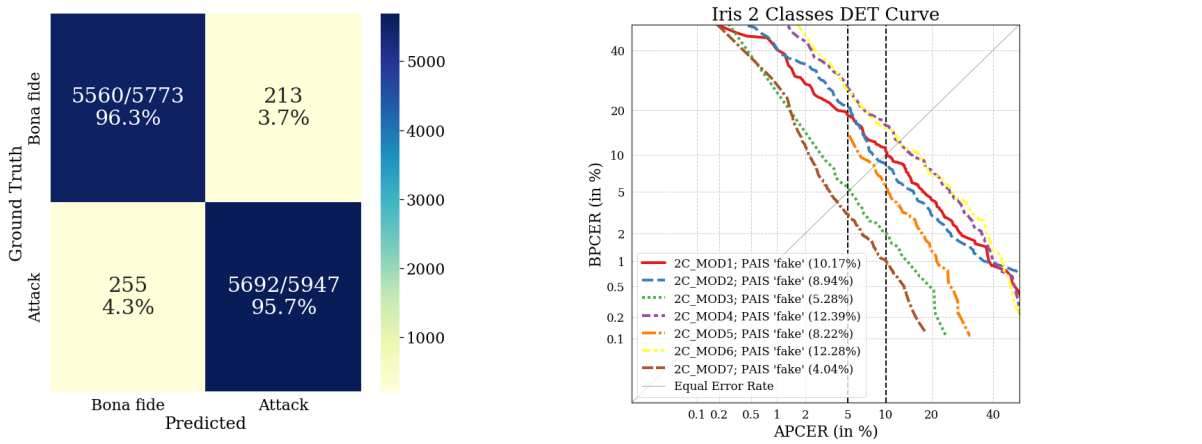


Fig. 4. Results of the PAD method with two classes for model 2C_MOD7. Left: confusion matrix for two class test, attack presentation (fake) and bona fide (live). Right: DET curve for the best results. The number in parenthesis corresponds to the EER in percentage. The black dashed lines indicate two operational points for $BPCER_{10}$ and $BPCER_{20}$.

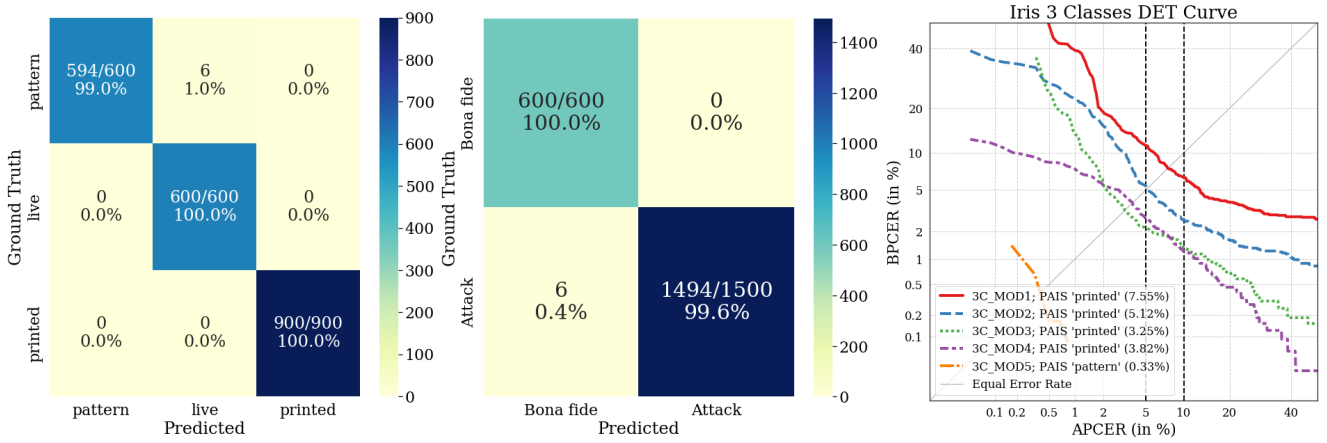


Fig. 5. Results of the PAD method with three classes for model 3C_MOD5. Left: confusion matrix considering each PAI independently. The second confusion matrix considers the bona fide class versus the fusion of all PAI species. The number in parenthesis corresponds to the EER in percentage. The black dashed lines indicate two operational points for $BPCER_{10}$ and $BPCER_{20}$.

VI. Experiments and Results

The approach presented in this work takes into account the variability of the attack presentation images, and the number of images per class. These images present a problem for the classifier because the PAI species are not equally represented (for instance only five images of cadaver eyes were available for LivDet-Iris 2020). Considering this imbalance, our strategy is primarily focused on classifying bona fide images with high precision first, and attack presentation images second. Therefore, our first approach was training a network with only two classes. Then, a second network was trained from scratch with three and four classes, increasing the number of filters (alpha 1.4) and weighting each class according to the numbers of images per species. To study these limitations and improve performance for these aforementioned scenarios, five experiments were developed in order to analyze the best hyper-parameter configuration of MobileNetV2. A combination of serial and parallel DNNs was used, trained from scratch. A grid search was used to determine the learning rate, number of epochs, global pooling operation, alpha value, and input size of images. All the experiments employ the CLAHE algorithm and the class weight balancing operation. All the networks were trained with a limit of 200 epochs, using an early stopping method in case the measured performance would stop improving. The input size of the image was 224×224 and 448×448 pixels. All the experiments used the same number of images.

A. Experiment 1

A traditional MobileNetv2 network was used, trained with fine-tuning techniques. Several tests were performed, sequentially freezing an additional MobileNetV2 block in each one, from the bottom of the network to the top. For this experiment the images were grouped in two classes: Bona fide and Fake. The Fake dataset encompasses all the different PAI species: Contact Lenses (CL), Printout (Pr), Electronic displays (EDs), Prosthetic Display (PD) and Cadaver Eyes (CE).

B. Experiment 2

A modified MobileNetv2a network was trained from scratch. For this experiment, the images were again grouped

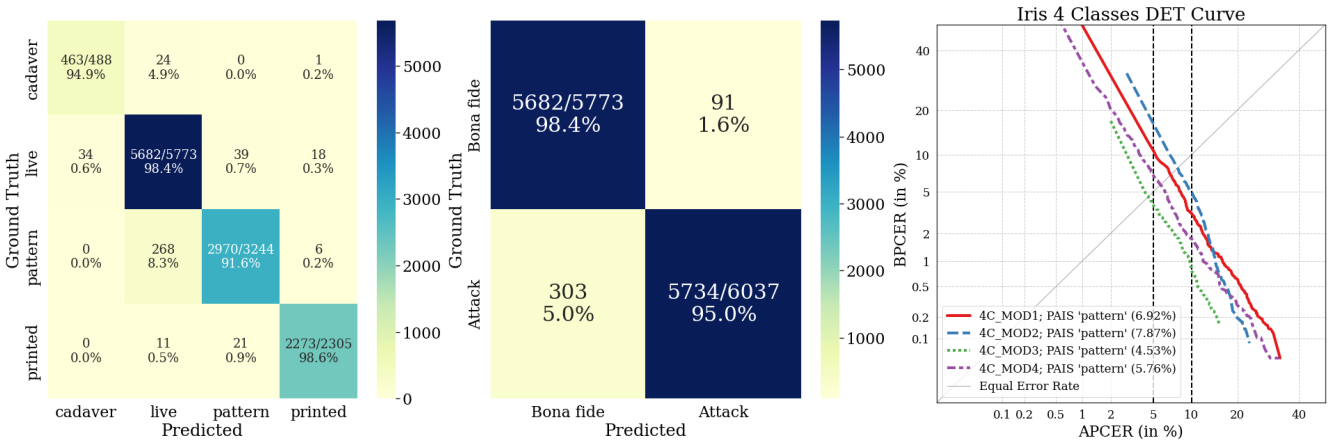


Fig. 6. Results of the PAD method with four classes for 4C_MOD3. Left: confusion matrix considering each PAI species independently. The second confusion matrix considers the bona fide class versus the fusion of all PAI species. The number in parenthesis corresponds to the EER in percentage. Similarly, The black dashed lines indicate two operational points for $BPCER_{10}$ and $BPCER_{20}$.

in two classes: BP (Bona fide presentations) and AP (Attack presentation with various PAI). The AP dataset is comprised of all PAI classes: Contact Lenses (CL), Printout (Pr), Electronic displays (EDs), Prosthetic Display (PD) and Cadaver Eyes (CE).

C. Experiment 3

For this experiment, a modified MobileNetv2b network was trained from scratch. The images were grouped in three classes this time: Bona fide, Contact lenses (patterned) and Printouts.

D. Experiment 4

A modified MobileNetv2b network were trained from scratch. The images in this experiment were grouped into four classes: Bona fide, Contact lenses, Printouts, and Cadaver.

E. Experiment 5

This experiment evaluates the feasibility of our proposed two-stage method against unknown/unseen PAI species, where these species are not part of the PAD algorithm training set. Three networks of two stages were trained, using a leave-one-PAI-species-out cross-validation protocol. The first model was trained using bona fide and printed images, and was evaluated using patterned contact lenses and cadaver PAI species. The second model was trained using bona fide and patterned contact lenses, and was evaluated using cadaver and printed PAI species. The last model was trained using bona fide and cadaver images, and was evaluated on patterned contact lenses and printed PAI species images.

F. Results

In this section, we report the best results for each experiment. Adam optimization performed better than SGD and RMSprop. The best initial learning rate was 1×10^{-5}. Global max pooling performs better than global average pooling. An alpha value of 1.0 performed better with two class scenarios, with an input image size of 224×224, whereas an

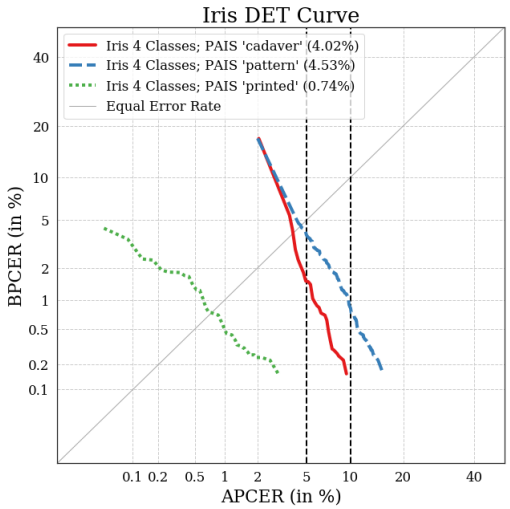


Fig. 7. DET curve for each PAI for the best four classes model.

alpha value of 1.4 with an input image size of 448×448 performed better for three and four class scenarios.

Table III shows an overview of the results for two class scenarios trained with fine-tuning, and two, three, and four class scenarios trained from scratch, which correspond to Experiments 1 to 4, respectively. Experiment 1 describes the fine-tuning approach. Experiments 2, 3 and 4 describe the two, three and four classes trained from-scratch respectively. For the Experiment 1, only the results for layers 10, 19, and 28 are included, due to the degradation in performance that was proportional to the amount of bottom layers from the network that were frozen. We infer this is probably because the pre-trained ImageNet [30] weights were not trained using images of spoof NIR eyes, or anything similar. Model names are IDs, which correspond to the curves shown in Figures 4 to 6. Overall, for fine-tuning, the best results were obtained when freezing only the first MobileNetV2 block (*2C_MOD7*), using Adam optimization, resulting in a $BPCER_{10}$ of 0.99% and $BPCER_{20}$ of 3.09%. Please note that for model *3C_MOD5*,

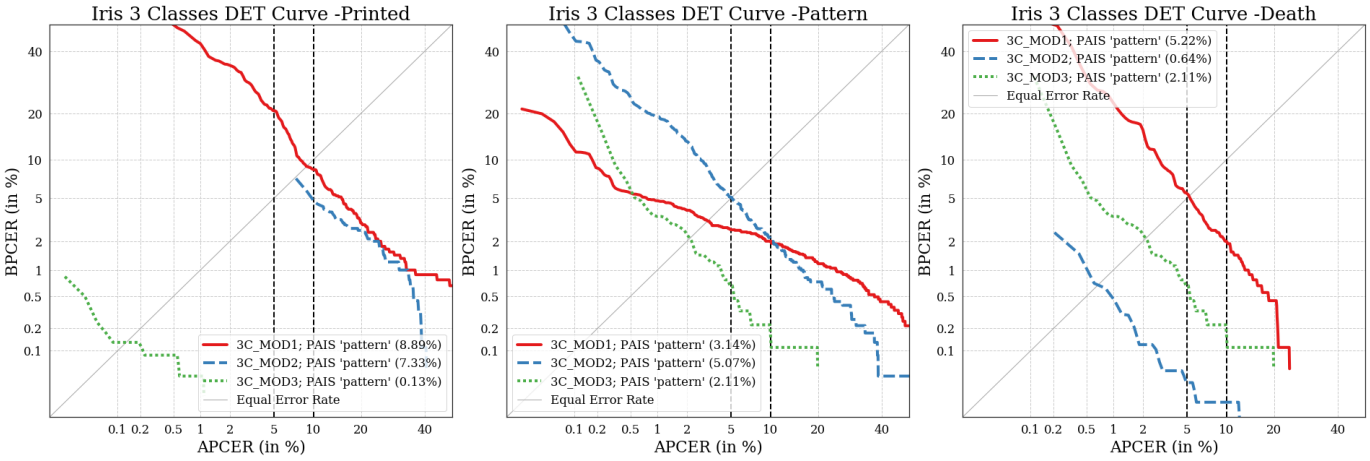


Fig. 8. DET curves for models trained using a leave-one-PAI-species-out cross-validation protocol. Models were trained using bona fide and printed PAI images, bona fide and patterned contact lenses PAI images, and lastly bona fide and post-mortem PAI images, respectively. Each model was evaluated using the PAI species that were not presented to them during training, effectively testing each model against unseen PAI species. The EER for each curve in percentage is shown in parenthesis. The black dashed lines indicate two operational points for $BPCER_{10}$ and $BPCER_{20}$.

TABLE III

SUMMARY OF THE RESULTS FOR TWO, THREE, AND FOUR CLASSES. IN BOLD ARE HIGHLIGHTED THE BEST RESULTS. POOL: GLOBAL POOLING OPERATION USED. FT: FINE TUNING TRAINING; NUMBER OF BLOCKS FROZEN. NETWORKS TRAINED FROM SCRATCH APPEAR AS "NONE". ACER, $BPCER_{10}$ AND $BPCER_{20}$ ARE REPORTED FOR ALL THE EXPERIMENTS. MODEL ID FORMAT IS "XC_MODY", WHERE "X" IS THE NUMBER OF CLASSES AND "Y" IS THE MODEL NUMBER

Model	ACER (%)	$BPCER_{10}$ (%)	$BPCER_{20}$ (%)	POOL	OPT	FT
2 CLASSES						
2C_MOD7	*3.99*	*0.99*	*3.09*	*AVG*	*ADAM*	*1st only*
2C_MOD6	9.50	15.33	25.66	AVG	SGD	1st & 2nd
2C_MOD5	6.51	5.66	15.12	AVG	ADAM	1st & 2nd
2C_MOD4	10.30	16.11	26.77	AVG	SGD	1st to 3rd
2C_MOD3	6.33	2.00	5.04	AVG	ADAM	1st to 3rd
2C_MOD2	5.44	8.52	20.72	AVG	SGD	NONE
2C_MOD1	6.55	10.55	18.44	AVG	SGD	NONE
3 CLASSES						
3C_MOD5	*0.5*	*0.16*	*0.16*	*AVG*	*ADAM*	*NONE*
3C_MOD4	1.70	3.13	3.30	AVG	ADAM	NONE
3C_MOD3	2.30	1.75	2.16	AVG	ADAM	NONE
3C_MOD2	2.31	3.71	5.34	AVG	ADAM	NONE
3C_MOD1	2.31	6.43	11.33	AVG	ADAM	NONE
4 CLASSES						
4C_MOD4	7.80	1.76	6.89	AVG	ADAM	NONE
4C_MOD3	*7.53*	*0.83*	*3.77*	*MAX*	*ADAM*	*NONE*
4C_MOD2	4.92	4.92	19.88	MAX	ADAM	NONE
4C_MOD1	5.38	3.16	10.73	MAX	SGD	NONE

the BPCER value corresponding to the highest APCER value is reported as $BPCER_{10}$ and $BPCER_{20}$, this is due that APCER values of 5% and 10% are never reached by this model, as can be seen in Figure 5.

Figure 4 shows the best result for two class scenarios, trained from scratch. This allows us to focus on identifying bona fide presentation (live) images versus attack presentation (fake) images. In this figure, a confusion matrix considering these two classes is shown. Additionally, a Detection Error Trade-off (DET) curve is presented. Several approaches were tested, where the best result reaches an EER of only 4.04% (brown curve), a $BPCER_{10}$ of 0.99%, and a $BPCER_{20}$ of 3.09%, respectively. The best model uses an alpha value of 1.4, an initial learning rate of 1×10^{-5}, and the Adam optimization algorithm.

Figure 5 shows the best result for three class scenarios: live, printed, and patterned contact lenses images. In this figure, a confusion matrix considering these three classes is shown. Furthermore, a confusion matrix showing bona fide presentation vs. attack presentation classes is presented. In this case, the attack presentation class encompasses both printed and patterned contact lenses PAI species. Additionally, a Detection Error Trade-off (DET) curve is also shown. The best result reaches an EER of only 0.33% (orange curve). For $BPCER_{10}$ and $BPCER_{20}$, a result of 0.16% is shown on Table III. This corresponds to the BPCER reached at the maximum possible APCER, which is 0.83%. As it can be seen in this DET curve, the 3C_MOD5 model never intersects with the black dashed lines. This model uses an alpha value of 1.4, an initial learning rate of 1×10^{-5}, and the Adam optimization algorithm.

Figure 6 shows the best result for four class scenarios: live, printed, patterned contact lenses, and post-mortem (cadaver). Likewise, two confusion matrices, one showing four classes, and the other grouping all PAI species under the "attack" class, are presented. A Detection Error Trade-off (DET) curve is also shown. The best result for this experiment reaches an EER of only 4.53% (green curve), a $BPCER_{10}$ of 0.83%, and a $BPCER_{20}$ of 3.77%. The best model uses an alpha value of 1.4, an initial learning rate of 1×10^{-5}, and the Adam optimization algorithm.

Figure 7 shows the performance for each PAI species for model 4C_MOD3. For the printed, cadaver, and pattern species Equal Error Rates of 0.74%, 4.02%, and 4.53% (as shown in Figure 6) were obtained, respectively.

TABLE IV

COMPARISON WITH THE STATE OF THE ART. RESULTS ARE SHOWN IN %. ALL THE METHODS WERE EVALUATED IN THE SAME TEST SET

Method	**APCER** (%)	**BPCER** (%)	**ACER** (%)
USACH/TOC (two-stages)	59.10	0.46	29.78
Fraunhofer-IGD	48.68	11.59	30.14
Competitor 3	57.8	40.31	49.06
NP PAD	57.21	0.71	28.96
MSU PAD Alg1	4.67	0.56	2.61
MSU PAD Alg2	2.76	1.61	2.18
DACNN	55.2	16.39	35.8
SIDPAD	49.85	39.96	44.9
Regional PAD	62.42	23.80	43.11

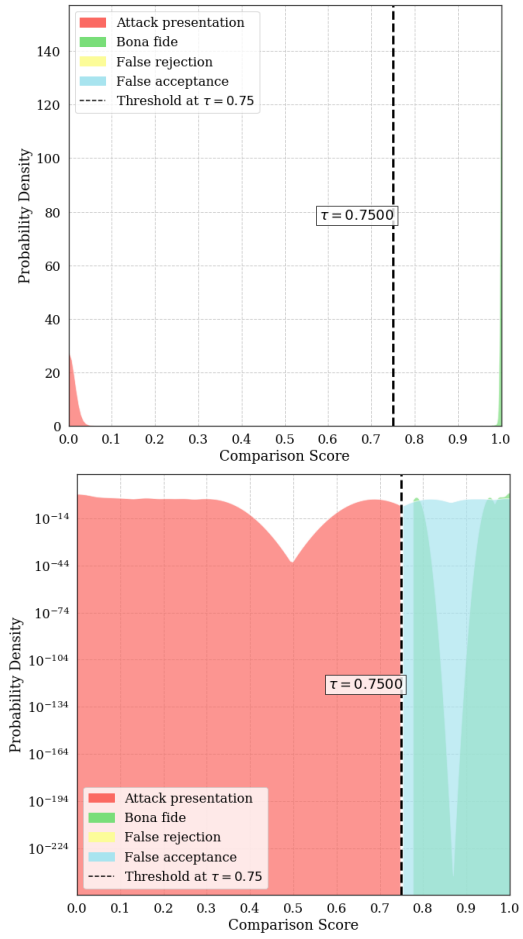


Fig. 9. Distribution of attack presentation scores versus bona fide scores. Top: linear scale. Bottom: logarithmic scale. The abscissa is shown in linear scale for both. The decision threshold is shown as the black dashed line.

Figure 8 shows the performance for unknown/unseen PAI species. Our proposal was evaluated using a leave-one-PAI-species-out cross-validation protocol. To that end, three two-stage networks were trained, according to Section VI-E. The Equal Error Rates reached are in the range of 2.11% to 8.89% for unseen PAI species. The best results show the robustness of our method in detecting unknown PAI species.

Finally, Table IV shows a comparison with the state-of-the-art methods, where our two-stage submitted proposal reached the best results on the LivDet-Iris 2020 Competition.

Additionally, Figure 9 shows two density plots—in linear and logarithmic scale for the ordinate respectively—showing the distribution of attack presentation scores versus bona fide scores for the best two-classes model 2C_MOD7, shown previously in Figure 4. The decision threshold is defined as 0.75 for demonstration purposes. This operating point can be adjusted depending on requirements and use case, subject to the trade-off between APCER and BPCER.

VII. CONCLUSION

Existing studies in the iris PAD literature are based on the assumption that the system encounters a specific iris presentation attack. However, this may not be the case in real-world scenarios, where the iris recognition system may have to handle multiple kinds of presentation attacks, including unseen species. We propose a framework focused on detecting bona fide images to address this challenge, which means optimising the models for a lower BPCER score. For this approach, we developed the largest iris presentation attack database by combining several other databases. This database is also available to other researchers by request.[4]

When trained from scratch, our suggested networks allow us to complement the results of the LivDet-Iris 2020 competition by using more challenging PAI species. When using fine-tuning, model performance worsens in proportion to the number of layers from the network that were frozen. Nonetheless, results using fine-tuning are competitive with the literature.

According to our results, an image input size of 224×224 is enough to classify bona fide images successfully. However, the results were improved for presentation attack instruments when using an image input size of 448×448. This result shows that the extra detail from higher resolution images contains relevant features for PAI species classification.

Overall the best result reached was with three scenarios, obtaining a BPCER_{10} of 0.16% and an EER of 0.33%.

This work reached competitive results according to the reported results in the LivDet-Iris 2020 competition.

For future work, newer lightweight model architectures such as MobileNetV3 [32], and EfficientNets [33] should be tested, and new PAI species should be included, considering, for example, synthetic images.

This work serves as the latest evaluation of iris PAD on a large spectrum of presentation attack instruments.

REFERENCES

[1] J. Tapia, C. Perez, and K. Bowyer, "Gender classification from the same iris code used for recognition," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 8, pp. 1760–1770, Aug. 2016.

[2] J. McGrath, K. W. Bowyer, and A. Czajka, "Open source presentation attack detection baseline for iris recognition," *CoRR*, vol. abs/1809.10172, pp. 1–8, Sep. 2018.

[3] J. Galbally, S. Marcel, and J. Fierrez, "Biometric antispoofing methods: A survey in face recognition," *IEEE Access*, vol. 2, pp. 1530–1552, 2014.

[4] S. Hoffman, R. Sharma, and A. Ross, "Convolutional neural networks for iris presentation attack detection: Toward cross-dataset and cross-sensor generalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 1620–1628.

[4] juan.tapia-farias@h-da.de

[5] D. Yambay *et al.*, "Iris liveness detection competition (LivDet-Iris)—The 2020 edition," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Sep. 2017, pp. 733–741.

[6] P. Das *et al.*, "Iris liveness detection competition (LivDet-Iris)—The 2020 edition," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Sep. 2020, pp. 1–9.

[7] M. Trokielewicz, A. Czajka, and P. Maciejewicz, "Post-mortem iris recognition with deep-learning-based image segmentation," *Image Vis. Comput.*, vol. 94, Feb. 2020, Art. no. 103866. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0262885619304597

[8] Z. Pan, W. Yu, X. Yi, A. Khan, F. Yuan, and Y. Zheng, "Recent progress on generative adversarial networks (GANs): A survey," *IEEE Access*, vol. 7, pp. 36322–36333, 2019.

[9] J. E. Tapia and C. Arellano "Soft-biometrics encoding conditional GAN for synthesis of NIR periocular images," *Future Gener. Comput. Syst.*, vol. 97, pp. 503–511, Aug. 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167739X18330966

[10] H. Zou, H. Zhang, X. Li, J. Liu, and Z. He, "Generation textured contact lenses iris images based on 4DCycle-GAN," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 3561–3566.

[11] Y. Hu, K. Sirlantzis, and G. Howells, "Iris liveness detection using regional features," *Pattern Recognit. Lett.*, vol. 82, pp. 242–250, Oct. 2016.

[12] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[13] E. Rahtu, J. Heikkilä, V. Ojansivu, and T. Ahonen, "Local phase quantization for blur-insensitive image analysis," *Image Vis. Comput.*, vol. 30, no. 8, pp. 501–512, Aug. 2012, doi: 10.1016/j.imavis.2012.04.001.

[14] D. Gragnaniello, G. Poggi, C. Sansone, and L. Verdoliva, "Using iris and sclera for detection and classification of contact lenses," *Pattern Recognit. Lett.*, vol. 82, pp. 251–257, Oct. 2016.

[15] D. Gragnaniello, C. Sansone, G. Poggi, and L. Verdoliva, "Biometric spoofing detection by a domain-aware convolutional neural network," in *Proc. 12th Int. Conf. Signal-Image Technol. Internet Syst. (SITIS)*, 2016, pp. 193–198.

[16] G. Y. Kimura, D. R. Lucio, A. S. Britto, Jr., and D. Menotti, "CNN hyperparameter tuning applied to iris liveness detection," 2020, *arXiv:2003.00833*.

[17] A. Boyd, A. Czajka, and K. Bowyer, "Deep learning-based feature extraction in iris recognition: Use existing models, fine-tune or train from scratch?" in *Proc. IEEE 10th Int. Conf. Biometrics Theory, Appl. Syst.*, Sep. 2020, pp. 1–9.

[18] D. Yadav, N. Kohli, A. Agarwal, M. Vatsa, R. Singh, and A. Noore, "Fusion of handcrafted and deep learning features for large-scale multiple iris presentation attack detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 572–579.

[19] D. Nguyen, N. Baek, T. Pham, and K. Park, "Presentation attack detection for iris recognition system using NIR camera sensor," *Sensors*, vol. 18, no. 5, p. 1315, Apr. 2018. [Online]. Available: https://europepmc.org/articles/PMC5981581

[20] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015, *arXiv:1409.1556*.

[21] 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 Trans. Inf. Forensics Security*, vol. 14, no. 6, pp. 1419–1431, Jun. 2019.

[22] D. Yambay, J. S. Doyle, K. W. Bowyer, A. Czajka, and S. Schuckers, "LivDet-iris 2013—Iris liveness detection competition 2013," in *Proc. IEEE Int. Joint Conf. Biometrics*, Sep. 2014, pp. 1–8.

[23] D. Yambay, B. Walczak, S. Schuckers, and A. Czajka, "LivDet-iris 2015—Iris liveness detection competition 2015," in *Proc. IEEE Int. Conf. Identity, Secur. Behav. Anal. (ISBA)*, Feb. 2017, pp. 1–6.

[24] A. Czajka and K. W. Bowyer, "Presentation attack detection for iris recognition: An assessment of the state-of-the-art," *ACM Comput. Surveys*, vol. 51, no. 4, pp. 1–35, 2018.

[25] A. B. Jung *et al.* (2020). *Imgaug*. Accessed: Feb. 1, 2020. [Online]. Available: https://github.com/aleju/imgaug

[26] S. Marcel, M. S. Nixon, J. Fiérrez, and N. W. D. Evans, Eds., *Handbook of Biometric Anti-Spoofing: Presentation Attack Detection* (Advances in Computer Vision and Pattern Recognition), 2nd ed. Cham, Switzerland: Springer, 2019, doi: 10.1007/978-3-319-92627-8.

[27] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.

[28] A. G. Howard *et al.*, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[30] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[32] A. Howard *et al.*, "Searching for MobileNetV3," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1314–1324.

[33] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Mach. Learn. Res.*, vol. 97, K. Chaudhuri and R. Salakhutdinov, Eds., Jun. 2019, pp. 6105–6114.

Juan E. Tapia (Member, IEEE) received the P.E. degree in electronics engineering from Universidad Mayor in 2004, the M.S. degree in electrical engineering from the Universidad de Chile in 2012, and the Ph.D. degree from the Department of Electrical Engineering, Universidad de Chile, in 2016. In addition, he spent one year of internship at the University of Notre Dame. In 2016, he received the award for best Ph.D. thesis. From 2016 to 2017, he was an Assistant Professor at Universidad Andres Bello. From 2018 to 2020, he was the Research and Development Director for the area of electricity and electronics at the Universidad Tecnologica de Chile. He is currently a Senior Researcher at Hochschule Darmstadt (HDA), and the Research and Development Director of TOC Biometrics. His main research interests include pattern recognition and deep learning applied to iris biometrics, morphing, feature fusion, and feature selection.

Sebastian Gonzalez (Member, IEEE) received the B.S. degree in computer engineering from Universidad Andres Bello in 2019. He is currently a Researcher at TOC Biometrics Company. His main interests include computer vision, pattern recognition, and deep learning applied to real problems, such as tampering detection, classification, and segmentation.

Christoph Busch (Member, IEEE) is currently a member of the Department of Information Security and Communication Technology (IIK), Norwegian University of Science and Technology (NTNU), Norway. He also holds a joint appointment with the Faculty of Computer Science, Hochschule Darmstadt (HDA), Germany. Further, he lectures the course biometric systems at Denmark's DTU since 2007. On behalf of the German BSI, he has been the Coordinator for the project series BioIS, BioFace, BioFinger, BioKeyS Pilot-DB, KBEinweg, and NFIQ2.0. In the European research program, he was an Initiator of the Integrated Project 3D-Face, FIDELITY, and iMARS. Further, he was/is a Partner in the projects TURBINE, BEST Network, ORIGINS, INGRESS, PIDaaS, SOTAMD, RESPECT, and TReSPAsS. He is also the Principal Investigator with the German National Research Center for Applied Cybersecurity (ATHENE). He coauthored more than 500 technical papers and has been a speaker at international conferences. Moreover, he is the Co-Founder and a Board Member of the European Association for Biometrics (www.eab.org) that was established in 2011 and assembles in the meantime more than 200 institutional members. He is a member of the Editorial Board of the *IET* journal.