IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 19 February 2024, accepted 30 April 2024, date of publication 6 May 2024, date of current version 21 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3397190

RESEARCH ARTICLE

Open-Set: ID Card Presentation Attack Detection Using Neural Style Transfer

REUBEN P. MARKHAM[1], JUAN M. ESPÍN LÓPEZ[2], MARIO NIETO-HIDALGO[2], AND JUAN E. TAPIA[3], (Member, IEEE)

[1]Instituto Tecnológico de Informática (ITI), Universitat Politècnica de València, 46022 Valencia, Spain
[2]Research and Development Centre, Facephi Biometria SA, 03003 Alicante, Spain
[3]da/sec-Biometrics and Internet Security Research Group, Hochschule Darmstadt, 64295 Darmstadt, Germany

Corresponding author: Juan E. Tapia (juan.tapia-farias@h-da.de)

This work was supported in part by the European Union (EU) Next-Generation, Plan de Recuperación, Transformación y Resiliencia through Convocatorias de Ayudas 2021, Proyecto Red.es, under Grant C005/21-ED; and in part by the German Federal Ministry of Education and Research and the Hessen State Ministry for Higher Education, Research and the Arts within their joint support of the National Research Center for Applied Cybersecurity (ATHENE).

ABSTRACT The accurate detection of ID card Presentation Attacks (PA) is becoming increasingly important due to the rising number of online/remote services that require the presentation of digital photographs of ID cards for digital onboarding or authentication. Furthermore, cybercriminals are continuously searching for innovative ways to fool authentication systems to gain unauthorized access to these services. Although advances in neural network design and training have pushed image classification to the state of the art, one of the main challenges faced by the development of fraud detection systems is the curation of representative datasets for training and evaluation. The handcrafted creation of representative presentation attack samples often requires expertise and is very time-consuming, thus an automatic process of obtaining high-quality data is highly desirable. This work explores ID card Presentation Attack Instruments (PAI) in order to improve the generation of samples with four Generative Adversarial Networks (GANs) based image translation models and analyses the effectiveness of the generated data for training fraud detection systems. Using open-source data, we show that synthetic attack presentations are an adequate complement for additional real attack presentations, where we obtain an EER performance increase of 0.63 % points for print attacks and a loss of 0.29 % for screen capture attacks.

INDEX TERMS Biometrics, synthetic images, remote verification, presentation attack detection, ID card.

I. INTRODUCTION

In recent years, a growing trend to digitalise processes that traditionally required physical attendance and the presentation of official ID documents has been observed. This tendency has been driven mostly by technological advances, novel legal regulations, and also pandemics, where long-term confinements force to look for alternatives to traditional ways of accessing certain services to remote services. Examples of affected processes are opening accounts in financial institutions, logistics, transport, retail or asset investment platforms, taking out insurance, and purchasing real estate [1].

The proliferation of remote services that require the identification of natural persons through biometrics and ID documents has motivated the continued search for weaknesses in the said process by the attacker in order to access the services without being identified. A common strategy is the presentation of documents that have been digitally manipulated and then printed on glossy or bond papers and represented on a smartphone or tablet screen, also known as spoofs or Presentation Attacks (PA).

The increasing sophistication and effectiveness of the methodologies with which attackers create convincing fake documents highlights the need to develop increasingly effective ID document Presentation Attack Detection (PAD). These systems have, as an essential component, an image classifier to distinguish between bona fide documents and

The associate editor coordinating the review of this manuscript and approving it for publication was Vincenzo Conti.

© 2024 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

PA. The current trend for creating image classifiers is to use neural network architectures, specifically Convolutional Neural Networks (CNN) [2] or Vision-Transformers (VT) [3], and to train them to minimize classification errors.

Moreover, it is well known that the stable training of modern neural networks requires a large set of diverse data to reach generalization capabilities. In the context of ID documents, data acquisition is a significant challenge because the data is subject to privacy concerns and legal regulations such as the GDPR,[1] which requires the consent of the subject for the processing and use of their data. Furthermore, the obtainment of PA would involve the laborious process of printing and cutting out of documents or preparing screen presentations with different display monitors [4]. The difficulties associated with the procurement of data have limited the quality and quantity of public research associated with the development of PAD models since the studies must often rely on in-house or private datasets, which makes it impossible to replicate the reported results (See Table 1). To alleviate these deficiencies and promote innovation, public datasets of synthetic documents have appeared in recent years, notably the Mobile Identity Document Video (MIDV) datasets [5], [6], [7], the Document Liveness Challenge (DLC-2021) dataset [8], and the Synthetic Chilean ID Card dataset [9].

This work leverages open-source datasets of video clips containing presentations of ID documents of fake subjects with the aim of ascertaining whether augmenting the training set with synthetic presentation attack samples instead of bona fide samples yields comparable results in terms of PAD predictive performance. To that end, two tasks are specified: firstly, the "print" task, where the model must distinguish between bona fide and print attack species, and lastly, the "screen" task, where bona fide and screen presentations are discriminated by the system. The datasets constructed for both tasks comprise of preprocessed frames of the original clips, where each image is a presentation of a full, aligned document with background information removed based on object detection systems [10], [11]. Thus, the three species considered in this work for classification are bona fide, print and screen:

- *Bona fide*: Video clips of ID cards containing synthetic data were captured in a variety of situations with smartphone cameras.
- *Print*: Digital templates of ID cards were printed on normal paper and cut out. Then, smartphones were used to capture short clips of the printed cards in different situations.
- *Screen*: Templates of ID cards were shown on computer and tablet screens, after which a smartphone was used to capture clips of the depicted images.

Supervised and unsupervised image-to-image translation models based on Generative Adversarial Networks (GANs) are explored to increase the number of presentation attack samples in the training dataset. This work is heavily inspired by a recent study [9] that employs GANs and texture transfer-based algorithms to generate bona fide and presentation attack samples. In the same vein as the aforementioned work, the usefulness of the generated images is assessed by training several MobileNetV2 [12] networks for each task. This way, the predictive performance using training sets comprised of synthetic and real samples can be compared with that of systems obtained by training with only real data.

In summary, the main contributions of this paper are:

- This work proposes a comprehensive analysis of the State-Of-The-Art related to PAD on ID cards and open-access databases.
- The GAN-based methods are explored in order to generate synthetic images to simulate and replicate the print and screen PA. This reduces the time to produce handcrafted attacks.
- Supervised and unsupervised presentation attack generation methods based on GANs are developed from supervised and unsupervised data for generating PA of full ID cards that retain the content of the original bona fide images.
- The system is trained using only open-access databases instead of private images used in the SOTA (Not available). Then, we show the improvements, limitations and tangible results we can reach with the open-access databases.
- This work shows and highlights the competitive results of developing an ID card PAD system based on open-access databases.

The remainder of the paper is structured as follows: Section II briefly discusses related works on GANs and ID cards PAD systems. Section III describes the methods used for generating new PA. The data and preprocessing used in the experiments are described in Section IV, while the metrics used to evaluate image quality and PAD predictive performance are presented in Section V. In Section VI, we detail the applied experimental framework and discuss the results. Finally, we provide a summary of our results in Section VII.

II. RELATED WORK

In the present section, we introduce the Neural style transfer concept and GAN models used to create synthetic presentation attack samples and briefly present fake ID detection systems found in recent literature.

Neural style transfer[2] is an optimization technique used to take two images—a content image and a style reference image (such as an artwork by a famous painter)—and blend them together so the output image looks like the content image, but "painted" in the style of the style reference image. This is implemented by optimizing the output image to match the content statistics of the content image and

[1] https://gdpr-info.eu/

[2] https://www.tensorflow.org/tutorials/generative/style_transfer

the style statistics of the style reference image. These statistics are extracted from the images using a convolutional network [13].

A. GENERATIVE MODELS

The traditional GAN [14] is a pair of neural networks that approaches the data generation task by implicitly modelling the distribution of a given dataset. It is composed of a generator G and a discriminator D network. The generator tries to generate data that is indistinguishable from the real data, whereas the discriminator tries to determine correctly whether a given data sample is real or fake. Both networks are trained simultaneously in a competitive manner, taking the form of a zero-sum game between two players where the objective is to find the Nash equilibrium.

With traditional GANs, there is scarce control on the output of G since it solely depends on the input noise vector. Conditional GANs were introduced in [15] that allow for greater control of the output by conditioning G and D on additional data $\mathbf{x}$ while training.

Both traditional and conditional GANs are the building blocks of the generative models used in this work. We focused on methods that approach the unimodal image-to-image translation task of finding a mapping G between input $\mathcal{X}$ and output $\mathcal{Y}$ image domains such that $\hat{\mathbf{y}} = G(\mathbf{x})$ is indistinguishable from the images of $\mathcal{Y}$. Four such methods are presented below: pix2pix [13], pix2pixHD [16], CycleGAN [17] and CUT [18]. The first two are supervised methods requiring pixel-aligned data for training, while the last two are unsupervised and trained on unpaired input and output image sets. The difference between paired and unpaired data is shown in Fig. 1.

pix2pix is based directly on the conditional GAN architecture. It uses as input to both networks the concatenation the original input and conditioning images on the channel dimension. In addition to the GAN loss, the authors use a $L1$ loss to enforce correctness at the low frequencies. A U-Net [19] architecture is used as the generator, while a novel "PatchGAN" architecture is used as the discriminator that classifies patches as real or fake and aggregates the results.

pix2pixHD aims to improve upon pix2pix for high-resolution image generation. The authors introduce a novel coarse-to-fine generator and a multi-scale discriminator architecture, as well as an improved adversarial loss based on matching the discriminator features at different layers. Furthermore, pix2pixHD allows conditioning with instance boundary maps and semantic label maps to improve the rendering of object boundaries.

Obtaining image pairs is relatively simple for certain tasks, such as superresolution, colourization, and inpainting. However, it can be prohibitive for a number of other tasks, such as translating photos into landscape paintings. This motivated the search for methods that could accomplish domain translation between unpaired sets of images.

CycleGAN achieves unpaired image translation by using two GANs and enforcing a cycle consistency loss between the generators. That is, given the generator G from $\mathcal{X}$ to $\mathcal{Y}$ and the generator F from $\mathcal{Y}$ to $\mathcal{X}$, the authors add to the GAN losses the following in Equation (1):

$$\begin{aligned}\mathcal{L}_{\text{cyc}}(G, F) = \mathbb{E}_{\mathbf{x}}[||F(G(\mathbf{x})) - \mathbf{x}||_1] \\ + \mathbb{E}_{\mathbf{y}}[||G(F(\mathbf{y})) - \mathbf{y}||_1]\end{aligned} \tag{1}$$

The cycle consistency condition, although an effective strategy to approach the unpaired translation problem, tends to force G into generating samples that contain all the necessary information in order to translate back to the input image, which leads to unsatisfactory results if significant visual changes are expected.

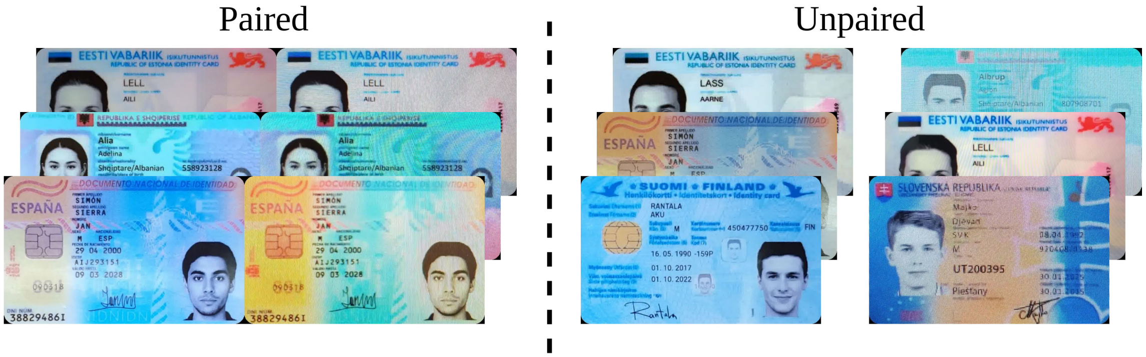
CUT uses a patch-wise contrastive loss [20] to maintain content correspondence between input and output images. The loss enforces similarity between corresponding patches of the input and generated images while enforcing dissimilarity with negative patches from the input image. The authors propose to use an encoder-decoder architecture for the generator, where the contrastive losses are computed on patches of features extracted from the encoder.

B. FAKE ID DETECTION

The widespread use of smartphones has prompted the development of novel remote authentication systems embedded in applications that require the input of biometric data such as fingerprints, faces, iris [21], and selfies [22]. Additionally, many services require digital photographs of ID cards, which are often captured with smartphones, as part of their digital onboarding process [23], [24]. Methods found in the literature that tackle the problem of remotely detecting fake ID cards from digital photographs are presented below.

Berenguel et al. [25] developed a novel application to detect ID documents that have been forged by a scan-printing operation. Their application allows the capture of Spanish ID documents using a mobile device and the assessment of their validity. The counterfeit detection module performs texture descriptor extraction, principal component analysis and feature pooling to classify regions of interest with linear Support Vector Machines (SVM). The final decision of labelling a document as genuine or counterfeit is performed by a naïve Bayes classifier. Additionally, Berenguel et al. [26] proposed a counterfeit document detector that uses a recurrent comparator architecture with attention models to spot the differences between a genuine and a reference image. The authors applied the detector to datasets of Spanish ID documents and banknotes. The system searches for the lack of resolution due to a scanning-printing operation by iteratively centring the attention on different positions of the security background textures and computing the differences.

Gonzalez et al. [1] presented a two-stage method for detecting tampered ID cards, which was trained and evaluated on a database of real Chilean national ID cards. The proposed


Paired

Unpaired

FIGURE 1. Paired training data (left) consists of pairs of pixel-aligned images. Unpaired data (right) comprises of two sets of images.

method uses a pre-trained MobileNet [27] to detect borders in the photo ID zone caused by composite tampering, while a second lightweight CNN, termed "BasicNet", was trained from scratch to detect the physical source of the document.

The DLC-2021 dataset Polevoy et al. [8] presented and defined three detection tasks: 1) screen recapture detection, where centre crops of documents from the original frames are classified as bona fide or screen recapture presentation, 2) unlaminated colour copy detection, where the network classifies scaled down images grey images as print presentation of bona fide, and 3) grey copy detection, where the classification is performed on projective undistorted document images. The authors train variations of the ResNet-50 [28] architecture on each task and report the results for future reference.

Mudgalgundurao et al. [29] proposed a pixel-wise supervision methodology which is used, along with a binary classification objective, to train presentation attack detectors on an in-house database of German ID cards and residence permits. The proposed system uses a simplified DenseNet [30] architecture, which the authors compare against baseline face PAD approaches.

Chen et al. [31] employed a scheme based on Siamese networks for document recapture detection. The network is trained on triplets of patches extracted from bona fide, recaptured, and reference documents. A custom "forensics loss" is used to attract genuine and reference representations while repelling recaptured and reference representations. The authenticity of a questioned document is evaluated using the distance metrics from three triplets. The authors created a database of synthetic university student ID cards to test their system.

Benalcazar et al. [9] explored the effectiveness of computer vision algorithms and generative models for the purpose of data augmentation while training fraud detection networks. The authors propose populating templates with synthetic data to create additional bona fide presentations, as well as training a StyleGAN-ADA [32] network to generate synthetic bona fide samples from scratch. For creating attack presentations, they use, in addition to the latter network, a texture transfer method based on adding artificial textures to bona fide presentations and a CycleGAN [17] model to translate between bona fide and attack domains. Various MobileNetV2 [12] models were trained on different combinations of real and synthetic Chilean national ID card presentations to assess the quality of the generated images. The authors report a negligible performance loss when supplementing databases with synthetic images.

TABLE 1. ID card dataset availability in the SOTA.

Author	ID card type	Open-Access
Berenguel et al. [25]	Spanish national	✗
Berenguel et al. [26]	Spanish national	✗
Gonzalez et al. [1]	Chilean national	✗
Polevoy et al. [8]	Various national	✓
Mudgalgundurao et al. [29]	German national and residence permits	✗
Chen et al. [31]	University student	✓
Benalcazar et al. [9]	Chilean national	✗
Magee et al. [33]	Brazilian national	✗

Magee et al. [33] explored the potential application of the Meijering filter [34] to the domain of recaptured identity document detection. The authors create a new dataset of recaptured images based on the publicly available BID [35] dataset and use it to train an SVM classifier on the raw histogram data obtained by using the filter. Although their system does not compare well with approaches that utilize neural networks, it remains an attractive alternative due to being transparent and explainable.

Most of the aforementioned studies train and test their proposed systems on private datasets using presentations of bona fide ID cards obtained from Gubernamental entities, company services, and banks in order to prevent fraud. As such, it is difficult to scrutinize and improve upon these systems since the data cannot be distributed publicly due to privacy concerns. In light of these challenges, some studies, though few in number, have created and published datasets composed of synthetic ID cards generated from templates, as seen in Table 1. However, they have limited commercial applicability because of the reduced number of subjects for each bona fide and attack image. These efforts are crucial for the effective public benchmarking of novel PAD systems.

III. METHODS

This section details the implementation of the GAN-based models used to generate additional presentation attack samples and of the PAD system used for both the screen and print tasks.

The proposed generative methods are designed to create synthetic PA of full ID cards. The composition of the datasets used to train each model is detailed in Section IV. The ID cards under study have widths of 85-86 mm and heights of 53-55 mm, while the employed image generator requires the input width and height to be a multiple of 8; thus an image size of 448 × 728 was chosen after preprocessing. The training of the generative models is performed on crops of these images, whereas the full images are used for inference.

A. IMAGE GENERATION FROM GANS

The GAN-based image-to-image translation methods used in this work to generate new samples of presentation attacks from bona fide presentations are pix2pix, pix2pixHD, CycleGAN and CUT. The first two are trained on paired data, while the remaining two simply require unpaired sets of images, one set per domain.

For the print task, the methods learn to transfer the visual characteristics of printed documents, such as the paper texture and fine-grain elements left by printers and ink. On the other hand, the pixel grid texture, spatial aliasing, and colour distortions indicative of screen displays are expected to be learned and faithfully transferred for the screening task.

An automatic procedure to generate pixel-aligned paired training data for pix2pix and pix2pixHD was implemented. Firstly, each bona fide image is paired randomly with a presentation attack image of the same subject. Next, the ORB algorithm [36] is used on each image to detect key points and extract binary local invariant features. Then, the Hamming distance between the features of one image and the features of the other is computed, and the best matches are found using an iterative algorithm. Finally, the homography matrix is estimated from the comparisons and is applied to align the presentation attack to the bona fide presentation.

For all the methods, we adopt the generator architecture from [37] with 9 residual blocks. Additionally, the 70 × 70 PatchGAN [13] architecture with 4 convolutional layers was used for the discriminator. We used three PatchGAN models for the multi-scale discriminator of pix2pixHD.

Training for each method and task was performed for 200 epochs with a batch size of 1 on random crops of size 224 × 224 × 3. Adam [38] was used as the optimizer with an initial learning rate of $2e-4$ and $\beta_1 = 0.5, \beta_2 = 0.999$. The learning rate for pix2pix was maintained fixed throughout training, while for the other methods, it declined linearly to zero after the 100th epoch. During inference, bona fide presentations of size 448 × 728 × 3 are fed to the generators to produce synthetic PA of the same size. The execution of training and inference was performed on a server with 32 CPU cores, 236 GB of RAM and a GPU of 40GB.

B. FRAUD-DETECTION NETWORKS

The same network architecture was used for both the screen and print tasks. Following [9], we used as the backbone a MobileNetV2 [12] pre-trained on ImageNet. The input to the networks is the 448 × 448 × 3 center crop of each image normalized with ImageNet mean and variance. The weights of the backbone are frozen during training. The output of the backbone is fed to a dropout layer [39] with $p = 0.2$, and the result is in turn fed to a final linear layer.

We use a batch size of 128 and train for 100 epochs. The weights are optimized using AdamW [40] with a constant learning rate of $5e-4$. Training and inference were performed on the same server as the generative models, with 32 CPU cores, 236 GB of RAM, and a GPU of 40GB.

IV. DATASETS

This section describes the datasets used in this work. Most of them present an important variability in light, illumination, background, orientation and others. It is essential to highlight that estimating the capture quality of ID cards is still an open problem [41].

Specifically, the source datasets are presented, and the print and screen tasks are formally defined (See Table 2. We use approximately 48,350 images derived from open-source datasets for our study. Furthermore, around 21,700 images of synthetic PA were generated to augment PAD model training sets.

Two experiments are defined:

- Experiment 1: the "print" task, where the PAD systems are meant to distinguish between bona fide and coloured print attack presentations.
- Experiment 2: the "screen" task, where bona fide and screen attack presentations are differentiated. The details of how the dataset for each task is constructed are provided below.

The list of images used to replicate and compare this proposal will be available for research purposes (upon paper acceptance).

A. DATASET ORGANIZATION

The MIDV-2020 and DLC-2021 datasets were used for this work.[3] Both are successors of MIDV-500 and consist of short video clips of fake documents presented in different lighting and background situations.

In MIDV-2020, the physical, bona fide documents were captured vertically in a resolution of $2,160 \times 3,840 \times 3$ pixels with 60 frames per second using a Samsung S10 or an Apple iPhone XR. In DLC-2021, physical and printed documents were captured, as well as screen presentations of the templates, where two office desktops and two notebook LCD monitors were used. The capturing was done with the

[3] MIDV-2020 is available to download from ftp://smartengines.com/midv-2020/dataset/. DLC-2021 is available in three parts from https://zenodo.org/records/7467028, https://zenodo.org/records/7467004 and https://zenodo.org/records/7467000

TABLE 2. Types of documents used in the experiments.

Document type code	Description	PRADO code
alb_id	Albanian ID document	ALB-BO-01001
esp_id	Spanish ID document	ESP-BO-03001
est_id	Estonian ID document	EST-BO-03001
fin_id	Finnish ID document	FIN-BO-06001
svk_id	Slovakian ID document	SVK-BO-05001

TABLE 3. Number of images per partition and class for each task.

Exp	Class	Train	Validation	Test
Exp1: Print	Bona fide	6,782	3,285	3,317
	Print	6,720	3,212	3,386
Exp2: Screen	Bona fide	4,066	3,224	3,633
	Screen	3,891	3,239	3,596

TABLE 4. Sets of bona fide and synthetic images used to create the experiment training sets.

Task	Class	$\mathcal{T}_A$	$\mathcal{T}_B$	$\mathcal{T}_B^b$	$\mathcal{T}_B^s$
Print	Bona fide	3,391	3,391	3,391	0
	Print	3,360	3,360	0	3,391
Screen	Bona fide	2,033	2,033	2,033	0
	Screen	1,945	1,946	0	2,033

same devices in two different frame resolutions ($1,080 \times 1,920 \times 3$ and $2,160 \times 3,840 \times 3$) and two different frame rates (30, 60 frames per second). In both datasets, the frames of the clips were extracted, and the authors annotated the position of the document.

The splits defined for the print and screen tasks were defined on the subject level. Each split contains at least two subjects for each type of document. Furthermore, bona fide representations from DLC-2021 of subjects with PA are included in the corresponding split. Additional subjects are then added from MIDV-2021 to ensure an approximately equal number of presentations for each class. In what follows, we specify which subjects from DLC-2021 and which from MIDV-2021 are included in each split for each task.

1) PRINT TASK

From DLC-2021, subjects 04-07 were used for the training data split, subjects 02 and 03 for validation data and subjects 00 and 01 for test data. From MIDV-2021, subjects 21-27 were used for the training data, subjects 32-38 for validation except for Albanian subject 35, and subjects 39-43 for test data.

2) SCREEN TASK

Subjects from DLC-2021 included in the training set are Albanian subjects 04 and 05, Spanish subjects 04, 05 and 06, Estonian subjects 04, 06 and 07, Finnish subjects 04, 05 and 07 and Slovakian subjects 04-07. The validation set contains DLC-2021 subjects 02 and 03, while the test set contains subjects 00 and 01. The training set contains no additional subjects.

From MIDV-2021, the validation includes Albanian subjects 17-22, Spanish, Estonian and Finnish subjects 18-23 and Slovakian subjects 19-24, while the additional subjects of the test set are Albanian subjects 09-15, Spanish, Estonian and Finnish subjects 10-16 and Slovakian subjects 11-17.

In both tasks, bona fide presentations come from both MIDV-2020 and DLC-2021, whereas attack presentations originate exclusively from DLC-2021. Moreover, only the data corresponding to ID cards, shown in Table 2, are considered, and images where the documents lie partially outside the frame are discarded.

The splits for each task are done on the subject level. At least two subjects must be present in each split and class. Furthermore, for the purpose of a fair comparison, we force the number of samples per class to be the same for each split.

Table 3, contains the number of images per split and class for the print and screen tasks. Approximately 13,500 images are assigned to the training set of the print task, followed by 6,500 images in the validation set and 6,700 in the test set. The train set of the screen task has 7,960 images, the validation set has 6,460 images, and the test set has 7,230 images.

Once the splits are created, the raw frames are preprocessed offline in three steps: firstly, the documents are projected to a $464 \times 744 \times 3$ rectangle via a perspective transformation using an estimated homography matrix; secondly, a portion of the background is masked out, and lastly, a centre crop of $448 \times 728 \times 3$ is applied. This process is illustrated in Fig. 2. Although the annotations were used to perform the projection in the first step, automatic segmentation of the document is possible with networks such as the one proposed by Lara et al. [42].

B. TRAINING SETS OF REAL AND SYNTHETIC IMAGES

The training set of each task is split into two disjoint sets, denoted by $\mathcal{T}_A$ and $\mathcal{T}_B$, in order to adequately assess the impact of using synthetic data. We denote by $\mathcal{T}_B^b$ the set of bona fide presentations of $\mathcal{T}_B$, which is used to generate the set of synthetic PA $\mathcal{T}_B^s$ for each method. Additionally, to reduce bias, we apply a mask to remove background information from generated images. The sizes of these datasets are shown in Table 4. Section VI describes how these sets are combined to create the training sets for the PAD systems so that the effect of synthetic data can be quantified.

V. METRICS

This section describes the metrics used, on the one hand, to evaluate the quality of generated images and, on the other, to assess the predictive performance of PAD systems.

A. QUALITY OF GENERATED IMAGES

Effective research concerning generative models relies upon metrics that can meaningfully assess the quality of generated images. In recent years many quantitative methods for computing the quality and diversity of synthetic data have been developed hand in hand with novel generative architectures, such as the Inception Distance [43], the Fréchet Inception Distance (FID) [44] and the Kernel Inception Distance [45].

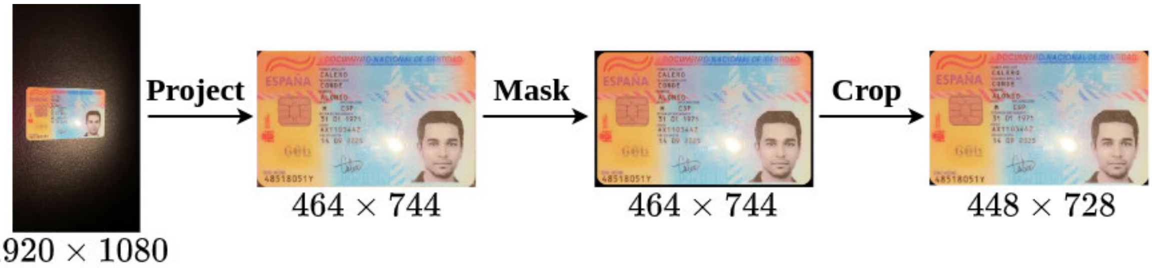


FIGURE 2. Preprocessing steps applied to the raw frame data.

The FID was selected for use in order to compare the results with the SOTA.

The FID is a similarity metric between two probability distributions that in practice, is often used to assess the similarity between two sets of images X and Y. To calculate the FID, firstly, the 2,048 dimensional feature vector (embedding) for each image is obtained by processing the image with a pre-trained Inception-v3 [46] network and keeping the features from the pool3 layer. Secondly, the per image set mean vectors μ_X, μ_Y and covariance matrices Σ_X, Σ_Y are calculated, and finally, the distributions are compared using the following Equation (2):

$$\mathrm{FID} = ||\mu_X - \mu_Y||_2^2 + \mathrm{Tr}\left(\Sigma_X + \Sigma_Y - 2(\Sigma_X \cdot \Sigma_Y)^{1/2}\right) \quad (2)$$

where $|| \cdot ||_2^2$ is the squared L^2 distance and $\mathrm{Tr}(\cdot)$ is the trace function.

B. DETECTION PERFORMANCE EVALUATION

The methodologies for evaluating the detection performance of biometric PAD algorithms are standardized by the ISO/IEC 30107-3[4] standard. The metrics used by this study are Attack Presentation Classification Error Rate (APCER), Bona fide Presentation Classification Error Rate (BPCER), $\mathrm{BPCER}_{\mathrm{AP}}$ and Equal Error Rate (EER). These metrics are aggregates of comparisons between the ground truth label $y \in \{0, 1, \ldots, J\}$ and the prediction $\hat{y}(\tau) \in \{0, 1, \ldots, J\}$ for a given operating point $\tau \in [0, 1]$, where $y = 0$ indicates a bona fide representation and $y = j$ with $j \geq 1$ is a presentation of the jth attack type.

To compute the APCER, firstly, the percentage of attack presentations incorrectly classified as bona fide is calculated for each presentation attack instrument as is shown in Equation (3):

$$\mathrm{APCER}_j(\tau) = \frac{100}{\sum_{i=1}^{N}[y_i = j]} \sum_{i=1}^{N} [y_i = j][\hat{y}_i(\tau) = 0] \quad (3)$$

where $[\cdot]$ is the Iverson bracket and N is the total number of presentations. Lastly, the maximum of these values, the worst-case scenario, is considered:

$$\mathrm{APCER}(\tau) = \max_j \mathrm{APCER}_j(\tau) \quad (4)$$

On the other hand, the BPCER is the percentage of bona fide presentations that have not been classified as is shown in Equation (5):

$$\mathrm{BPCER}(\tau) = \frac{100}{\sum_{i=1}^{N}[y_i = 0]} \sum_{i=1}^{N} [y_i = 0][\hat{y}_i(\tau) \neq 0] \quad (5)$$

The remaining two metrics analyze the system performance on specific operating points. The $\mathrm{BPCER}_{\mathrm{AP}}$ is the BPCER value when the APCER is fixed at 100/AP. In this work we evaluate BPCER_{10}, BPCER_{20} and BPCER_{100}, which correspond to APCER values of 10%, 5% and 1% respectively. The EER is the operating point where APCER = BPCER; however, the classification rate is often reported instead. In practice, there may not exist an operating point that satisfies the previous condition, thus, a reasonable interpolated value is often used.

The aforementioned metrics can be represented using Detection Error Tradeoff (DET) curves [47], which are also used by this study. They represent the APCER on the X axis and the BPCER on the Y axis and use a normal deviate scale for both axes, which spreads out the plot and facilitates the visual comparison of different systems.

VI. EXPERIMENTS AND RESULTS

This section describes the experiments performed on the data described in Section IV and reports the results using the metrics shown in Section V. First, we show how generative methods perform in terms of the visual similarity of generated samples with real data. Then, for each task, we analyze the effect on PAD predictive performance of adding synthetic samples to the training dataset instead of bona fide samples, which, in practice, are harder to obtain.

A. SYNTHETIC IMAGE QUALITY EVALUATION

The aim of this experiment is to assess the quality of synthesized PA by comparing them to sets of real PA. The comparison between the two sets of images is done with the FID metric. Each generative system was trained on $\mathcal{T}_A$ and applied on $\mathcal{T}_B^b$ to obtain the set of synthesized presentations $\mathcal{T}_B^s$. Then, the FID of $\mathcal{T}_B^s$ with the test presentation attack images is computed. This allows for the comparison of generative methods in terms of the visual quality of the generated samples. Additionally, the obtained FID values are compared to our baseline FID values obtained from the

[4] https://www.iso.org/standard/79520.html

TABLE 5. FID scores computed between synthetic fake images with proposed generation methods and test data.

Task	Validation	CycleGAN	CUT	pix2pix	pix2pixHD
Print	43.24	**55.73**	56.69	60.00	65.02
Screen	58.22	**68.77**	77.39	78.92	76.16

validation set in order to compare the difference between synthetic images and PA (screen, print).

Table 5, shows the results in terms of the FID scores. The second column, which reports the best results, shows the baseline FID scores. With a FID score of 43.24 for the print task and of 58.22 for the screen task, these values represent the ideal performance for a generative model. The best performing FID computed on synthetic data is shown in bold for each task, where CycleGAN obtains the best results for both tasks with FID scores of 55.73 and 68.77 for the print and screen task respectively.

In regards to synthesized print attacks, CycleGAN is closely followed by CUT, while pix2pix and pix2pixHD are the worst-performing methods with FID scores above 60. On the other hand, in terms of synthesized screen attacks, the next best method is pix2pixHD, closely followed by CUT and lastly pix2pix.

For both types of attack, we observe that unsupervised methods (CycleGAN & CUT) perform on par or better than supervised methods. We hypothesize that this is due in part to deficiencies in the automatic image alignment process to produce paired training data, which results in noisier images. Additionally, we suspect that the unnaturalness of the reconstructions of external elements in the source image, such as fingers or reflected light, could contribute to a higher FID score. These effects can be observed in Fig. 4, in which bona fide presentations of each document type, as well as the corresponding synthetic PA generated with each method, are presented.

The examples generated with CycleGAN shown in Fig. 4 preserve better the content of the bona fide presentation compared with samples of the same attack type generated with other methods. This is likely a result of the increased capacity of the cycle consistency loss for conserving information between translations. Moreover, in addition, the aforementioned problems with the supervised data, the increased complexity of the discriminator, and the low resolution and diversity of the training data are likely causes of the increased noise and artefacts observed in the samples obtained with pix2pixHD.

Additionally, from Fig. 3, it can be seen that there are noticeable differences in colour and brightness between the synthetic print presentations of each bona fide document. This may be due to a number of factors, such as the variability of lighting conditions, the presence of different document types in the dataset or the differences between the methods of preserving input colour information. However, the colour saturation of the generated samples appears to be lower than that of the corresponding bona fide presentation, which is expected when printing on matte or uncoated papers. The differences in colour and brightness between synthesized screen presentations seem to be less pronounced for certain documents, which can be attributed to the reduced variability of these aspects in screen displays. Moreover, some synthetic samples present a grid-like texture, which shows that the generative models have successfully learned to transfer this feature of screen displays. On the other hand, there is an absence of sophisticated moiré patterns in the generated samples, which is likely due to the small number of training samples with such patterns, as well as the distortion of the original patterns due to the projection of the document segments in the preprocessing stage.

B. EXPERIMENT 1: PAD PERFORMANCE ON PRINT TASK

For the print task experiments, the MobileNetV2 networks are binary image classifiers that detect whether the input presentation is bona fide or a print attack. In total, 6 networks were trained. The first network was trained using only $\mathcal{T}_A$ (6,751 images) in order to gauge the effect of adding more data. The second network was trained on $\mathcal{T}_A \cup \mathcal{T}_B$ (13,502 images) and represents the model trained with the complete set of real data. The remaining networks were trained on $\mathcal{T}_A \cup \mathcal{T}_B^b \cup \mathcal{T}_B^s$ (13,533 images), where $\mathcal{T}_B^s$ is different for each method, and represent the cases where synthetic data is used. The validation set of Table 3 was used to determine the best checkpoint, and the PAD metrics were calculated on the test set.

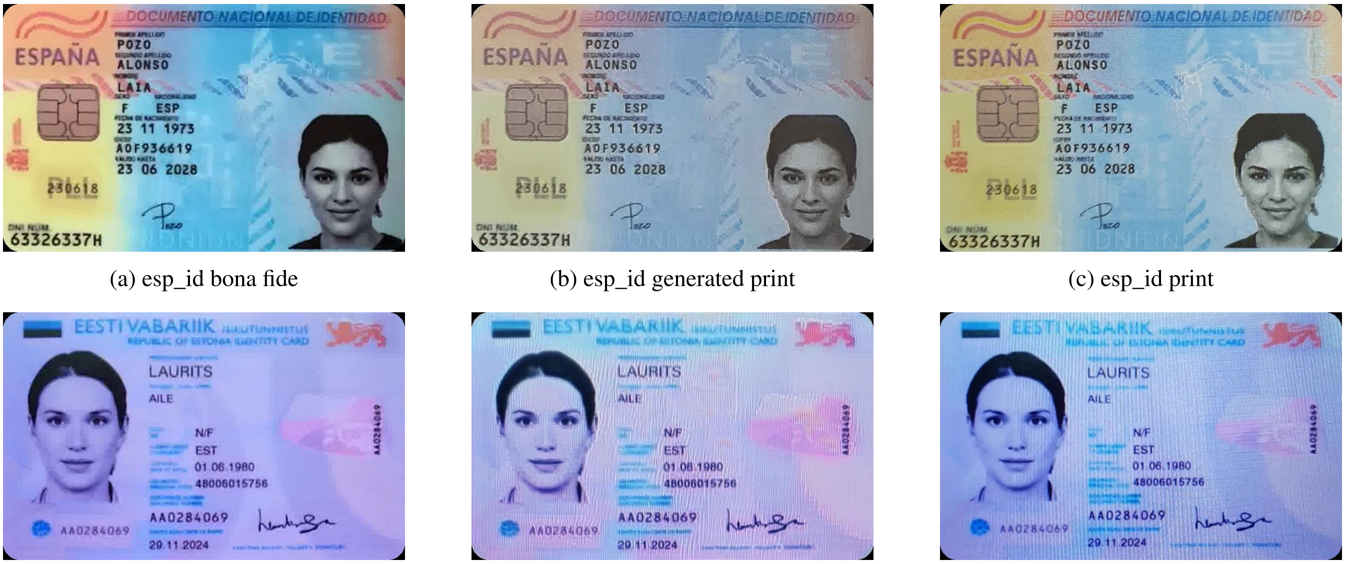
The DET curves obtained from predictions of the print task test set of all networks are displayed in Fig. 5, where the EER values for each curve are also reported. Interestingly, the best performance is obtained with synthetic PA generated with pix2pixHD, with an EER of 3.16%. With pix2pix data, we observe a slight drop in performance with an EER of 3.33%, but still better than using real PA were the value of 3.79% was observed. With CycleGAN data, we obtain an EER of 3.82%, which is comparable to using real data. CUT produced the worst performing data with an EER of 4.31%, which is slightly above the 4.28% obtained by only using $\mathcal{T}_A$.

Table 6 contains the BPCER$_{10}$, BPCER$_{20}$ and BPCER$_{100}$ operational points of each experiment. We observe a similar trend as reported with the EER values, except for BPCER$_{10}$ where pix2pixHD performs on par with real data with a value of 0.72% and all cases perform better than training only with $\mathcal{T}_A$ where a value of 1.90% was observed.

Given the previous observations, it can be said that data synthesized using supervised generative models produced better-performing PAD models than data generated with unsupervised methods. However, we observed in Section VI-A that supervised methods produced the data most dissimilar to real data. Hence, when synthetic print attacks are involved, the FID score correlates positively with PAD predictive performance.

C. EXPERIMENT 2: PAD PERFORMANCE ON-SCREEN TASK

The screen task models are trained to distinguish between bona fide and screen presentations of ID cards. We configured



(a) esp_id bona fide

(b) esp_id generated print

(c) esp_id print

(d) est_id bona fide

(e) est_id generated screen

(f) est_id screen

FIGURE 3. Examples of ID cards. (a) to (c) showcase examples of bona fide, generated print presentations and handcrafted print presentations of Spanish ID cards. (d) to (f) showcase examples of bona fide, generated screen presentations and handcrafted screen presentations of Estonian ID cards. (b) and (e) were generated with CycleGAN.

TABLE 6. Results for PAD models trained on the print task. Values expressed in %.

	$\mathcal{T}_A$	$\mathcal{T}_A \cup \mathcal{T}_B$	CycleGAN	CUT	pix2pix	pix2pixHD
EER	4.28	3.79	3.82	4.31	3.33	**3.16**
BPCER$_{10}$	1.69	0.72	1.03	1.39	0.96	**0.72**
BPCER$_{20}$	3.80	2.71	3.04	3.83	2.32	**2.17**
BPCER$_{100}$	10.16	9.35	9.20	12.63	8.20	**7.60**

TABLE 7. Results for PAD models trained on the screen task. Values expressed in %.

	$\mathcal{T}_A$	$\mathcal{T}_A \cup \mathcal{T}_B$	CycleGAN	CUT	pix2pix	pix2pixHD
EER	6.53	**5.80**	6.09	6.28	7.39	7.07
BPCER$_{10}$	1.90	**1.62**	2.67	2.39	3.72	3.47
BPCER$_{20}$	9.50	**7.10**	7.90	8.75	12.61	11.89
BPCER$_{100}$	43.24	32.56	**29.64**	40.77	38.67	42.42

the experiments in a similar manner to those of the print task, where 6 networks are trained using different combinations of the screen task datasets described in Table 4. The first evaluation uses $\mathcal{T}_A$ (3,978 images) as a training set, while the second uses the complete training set $\mathcal{T}_A \cup \mathcal{T}_B$ (7,957 images) comprised only of real data. The remaining experiments combine the first half of the training set with the bona fide images of the second half and the synthetic images generated from said bona fide images, that is $\mathcal{T}_A \cup \mathcal{T}_B^b \cup \mathcal{T}_B^s$ (8,044 images). After training, the best-scoring network checkpoint on the validation set is retained and evaluated on the test set of Table 3.

The DET curve of each screen task experiment is shown in Fig. 6 along with the corresponding EER score. The full training set of real data produces the best model with an EER of 5.80%. The next best model was trained with CycleGAN data with a score of 6.09%, followed by the one trained with CUT data with a score of 6.28%. The aforementioned models have better predictive performance in terms of EER score than the model trained only with $\mathcal{T}_A$, where a value of 6.53% was observed. The worst-performing models were trained on data generated with pix2pixHD and pix2pix, with EER scores of 7.07% and 7.39%, respectively.

A different trend can be observed from the BPCER$_{\text{AP}}$ values, reported in Table 7. The model trained on $\mathcal{T}_A \cup \mathcal{T}_B$ obtained the best BPCER$_{10}$ and BPCER$_{20}$ scores, with values of 1.62% and 7.10% respectively, while the model trained with CycleGAN data obtained the best BPCER$_{100}$ score with a value of 29.64%. Furthermore, all models trained with synthetic data obtained worse BPCER$_{10}$ scores than the model trained solely on $\mathcal{T}_A$ where a value of 1.90% was observed. BPCER$_{20}$ scores show a similar trend as the EER scores. With respect to the BPCER$_{100}$ scores, the worst performing model was trained on $\mathcal{T}_A$ where a value of 43.24% was observed, and pix2pixHD provided the worse performing synthetic training data with an observed value of 42.42%.

The results presented above show that the best-performing synthetic data was provided by unsupervised generative models, with CycleGAN performing better in general than CUT. Moreover, pix2pixHD performed better than pix2pix in three out of four metrics. This follows the same trend as seen with the FID scores seen in Table 5. We postulate that the noise introduced by the automatic pixel-alignment process hinders the performance of PAD models because they rely on high-level texture details to detect screen presentations adequately. A closer look at the presentations of the last four rows of Fig. 4 shows how the unsupervised models produce samples with greater detail and finer texture than supervised models. Besides, in some samples generated by the latter models, the alphanumeric information appears distorted, which might further degrade predictive performance.

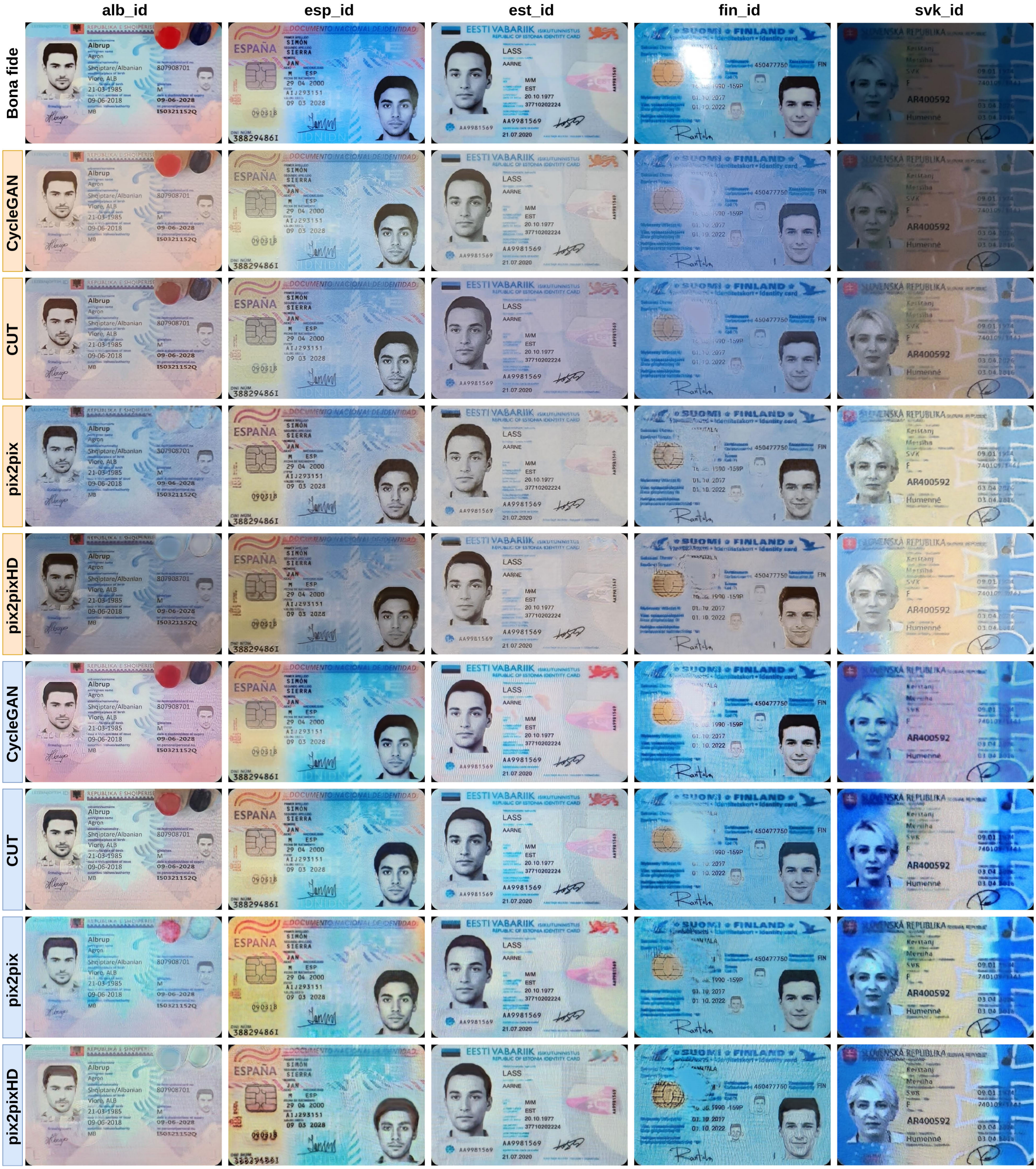


FIGURE 4. Examples of bona fide and synthetic images generated with the proposed methods. The first row contains bona fide samples of each type of ID card, and the following rows contain the output produced by each method. The method names with orange backgrounds were trained on the print task, while the ones with blue backgrounds were trained on the screen task.

D. DISCUSSION

The main focus of this article is to analyze the potential of GAN-based methods for generating effective PA of front-faced, full ID cards on open-access datasets. For this purpose, we first compared the generated data with real data using the FID metric, and then we evaluated their contribution to PAD predictive performance when replacing real data. However, the reported results are not directly comparable to those

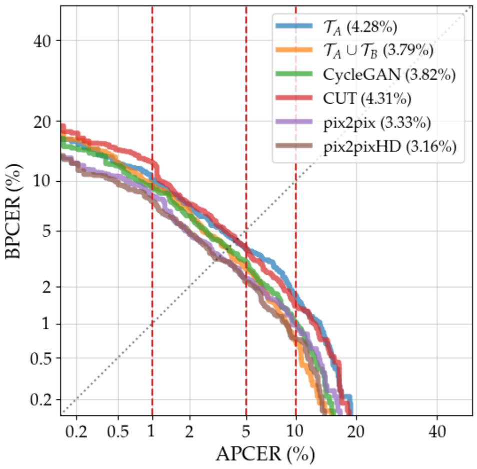


FIGURE 5. Detection Error Trade-off curves of networks trained on print task data. The EER is shown in parentheses for each scenario. Red dot lines represent three operational points BPCER_{10} and BPCER_{20} and BPCER_{100}, respectively.

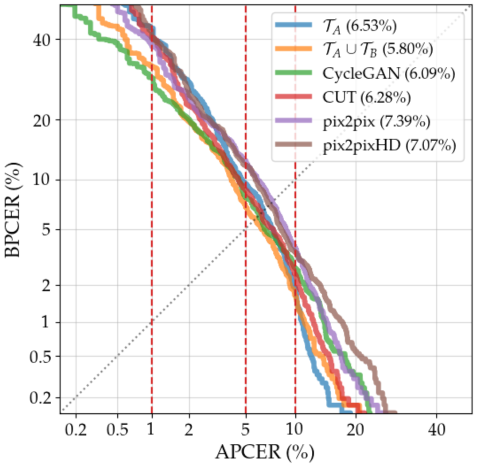


FIGURE 6. Detection Error Trade-off curves of networks trained on-screen task data. The EER is shown in parentheses for each scenario. Red dot lines represent three operational points BPCER_{10} and BPCER_{20} and BPCER_{100}, respectively.

of SOTA. Firstly, most of the papers in the SOTA using a private dataset are not available. Second, most studies used different types of ID cards (countries) than those used here and with different subject numbers. Third, all cited studies train their respective PAD models using non-projected data. Fourth, some studies use datasets comprised of more than one type of attack for training, as is the case in Benalcazar et al. [9]. Lastly, some studies, notably Polevoy et al. [8], fail to report predictive performance using ISO/IEC 30107-3 standardized metrics. In summary, about the limitations of the proposed methods, it is well known that the training of GANs is complex and often unstable Salimans et al. [43]. Moreover, training can lead to mode collapse, which is made evident by the presence of noticeable artefacts or the lack of variety in the output images, thereby negatively impacting the style transfer process. Additionally, pix2pix and pix2pixHD require a large set of aligned pairs of images to train, which is a challenge to obtain for PAD because of privacy concerns. On the other hand, CycleGAN and CUT do not require aligned pairs but lack the capability to remove external elements. Lastly, the proposed systems look for a trade-off capacity to transfer complex stylistic elements, such as Moiré patterns, between ID cards.

VII. CONCLUSION

In this work, we addressed the problem of open-access PA data scarcity by proposing methods that use GAN-based generative models to create synthetic samples. Additionally, we studied whether the generated data are an effective substitute for real data for training PAD models. For this purpose, we leveraged two open-source datasets containing ID card presentations of fake subjects.

We defined two experiments based on these datasets: the ‘‘print’’ task to distinguish between bona fide and print presentations and the ‘‘screen’’ task to differentiate between bona fide and screen presentations. We trained a total of 12 MobileNetV2 networks, 6 for each task, using different combinations of datasets comprised of real and generated data in order to adequately assess the impact of adding more training data and replacing real PA with synthesized ones.

The results vary greatly from one task to the other. Regarding the print task, data generated with the supervised generative models proved as effective or even more so than using additional real data, obtaining a 0.63% increase in performance with pix2pixHD data. On the other hand, the best-performing unsupervised model data proved as effective as additional real data, while the worst performance was achieved with CUT data, which is on par with not adding any data to the training set. With reference to the screening task, we observed that data synthesized with unsupervised methods proved slightly less effective than supplementary real data while still having a 0.44% advantage over not using additional data. On the contrary, data generated with pix2pixHD and pix2pix was detrimental to model performance, with a performance degradation of at least 0.54 % compared to using no additional data.

In all cases, we observed that CycleGAN data performs better than CUT data. As such, the cycle consistency mechanism for preserving content and transferring style is shown to be better suited for presentation attack generation than the patch-wise contrastive loss used by CUT.

On the other hand, we also observe that pix2pixHD data is more effective than pix2pix data for PAD predictive performance despite producing less visually appealing presentations. This may be due to the distortions' positive regularising effect, given the low variability in the training data, although further research is needed to validate these claims.

We also analyzed the quality of the generated images using the FID metric. The results reveal that CycleGAN produces the most faithful images, followed by CUT, pix2pixHD and finally pix2pix. This stands in contrast to the results observed in the print task experiments, where unsupervised models produced less effective data than supervised models. However, increased image quality is observed to be aligned with the screen task results. We attribute these differences in part to shortcomings in the image alignment process used for creating paired samples for supervised generative model training, which results in noisy generated samples. We hypothesize that this noise correlates well with the noise inherent to the paper texture while proving harmful for generating screen samples since the alignment process can interfere with the fine grain texture and moiré patterns expected in screen displays.

This article focused primarily on GAN-based unimodal, two-domain image-to-image translation models. Future work includes exploring the effectiveness of multi-modal models, as well as models that generate images in more than one domain. Additionally, we plan to complement our work by analyzing the quality of ID cards generated with recently proposed diffusion models.

ACKNOWLEDGMENT

The authors are grateful to Instituto Tecnológico de Informática.

REFERENCES

[1] S. Gonzalez, A. Valenzuela, and J. Tapia, "Hybrid two-stage architecture for tampering detection of chipless ID cards," *IEEE Trans. Biometrics, Behav., Identity Sci.*, vol. 3, no. 1, pp. 89–100, Jan. 2021.

[2] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Comput.*, vol. 1, no. 4, pp. 541–551, Dec. 1989.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. U. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2017, pp. 5998–6008. [Online]. Available: https://papers.nips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Metadata.json

[4] J. Tapia, C. Busch, H. Zhang, R. Ramachandra, and K. Raja, "Simulating print/scan textures for morphing attack detection," in *Proc. 31st Eur. Signal Process. Conf. (EUSIPCO)*, Sep. 2023, pp. 610–614.

[5] V. V. Arlazarov, K. Bulatov, T. Chernov, and V. L. Arlazarov, "MIDV-500: A dataset for identity document analysis and recognition on mobile devices in video stream," *Comput. Opt.*, vol. 43, no. 5, Oct. 2019. [Online]. Available: http://computeroptics.ru/eng/KO/Annot/KO43-5/430515e.html

[6] K. Bulatov, D. Matalov, and V. V. Arlazarov, "MIDV-2019: Challenges of the modern mobile-based document OCR," in *Proc. 12th Int. Conf. Mach. Vis. (ICMV)*, Jan. 2020, pp. 717–722.

[7] K. Bulatov, E. Emelianova, D. Tropin, N. Skoryukina, Y. Chernyshova, A. Sheshkus, S. Usilin, Z. Ming, J.-C. Burie, M. Muzzamil Luqman, and V. V. Arlazarov, "MIDV-2020: A comprehensive benchmark dataset for identity document analysis," 2021, *arXiv:2107.00396*.

[8] D. V. Polevoy, I. V. Sigareva, D. M. Ershova, V. V. Arlazarov, D. P. Nikolaev, Z. Ming, M. M. Luqman, and J.-C. Burie, "Document liveness challenge dataset (DLC-2021)," *J. Imag.*, vol. 8, no. 7, p. 181, Jun. 2022. [Online]. Available: https://www.mdpi.com/2313-433X/8/7/181

[9] D. Benalcazar, J. E. Tapia, S. Gonzalez, and C. Busch, "Synthetic ID card image generation for improving presentation attack detection," *IEEE Trans. Inf. Forensics Security*, vol. 18, pp. 1814–1824, 2023.

[10] L. Li, X. Yao, X. Wang, D. Hong, G. Cheng, and J. Han, "Robust few-shot aerial image object detection via unbiased proposals filtration," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5617011. [Online]. Available: https://ieeexplore.ieee.org/document/10197537

[11] X. Qian, B. Wu, G. Cheng, X. Yao, W. Wang, and J. Han, "Building a bridge of bounding box regression between oriented and horizontal object detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5605209. [Online]. Available: https://ieeexplore.ieee.org/document/10068217

[12] 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.

[13] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5967–5976.

[14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, Eds. Red Hook, NY, USA: Curran Associates, 2014, pp. 2672–2680. [Online]. Available: https://papers.nips.cc/paper_files/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Metadata.json

[15] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*.

[16] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional GANs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8798–8807.

[17] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2242–2251. [Online]. Available: http://ieeexplore.ieee.org/document/8237506/

[18] T. Park, A. A. Efros, R. Zhang, and J.-Y. Zhu, "Contrastive learning for unpaired image-to-image translation," in *Computer Vision—ECCV*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham, Switzerland: Springer, 2020, pp. 319–345.

[19] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, vol. 9351. Cham, Switzerland: Springer, Oct. 2015, pp. 234–241.

[20] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748*.

[21] J. E. Tapia, S. Gonzalez, and C. Busch, "Iris liveness detection using a cascade of dedicated deep learning networks," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 42–52, 2022.

[22] J. E. Tapia, A. Valenzuela, R. Lara, M. Gomez-Barrero, and C. Busch, "Selfie periocular verification using an efficient super-resolution approach," *IEEE Access*, vol. 10, pp. 67573–67589, 2022.

[23] H. Wang, S. Li, S. Cao, R. Yang, J. Zeng, Z. Qian, and X. Zhang, "On physically occluded fake identity document detection," in *Proc. 31st ACM Int. Conf. Multimedia*. New York, NY, USA: Association for Computing Machinery, Oct. 2023, p. 1556.

[24] J. Li, C. Kong, S. Wang, and H. Li, "Two-branch multi-scale deep neural network for generalized document recapture attack detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.

[25] A. Berenguel, O. R. Terrades, J. Lladós, and C. Cañero, "E-counterfeit: A mobile-server platform for document counterfeit detection," in *Proc. 14th IAPR Int. Conf. Document Anal. Recognit. (ICDAR)*, vol. 9, Nov. 2017, pp. 15–20.

[26] A. B. Centeno, O. R. Terrades, J. L. Canet, and C. C. Morales, "Recurrent comparator with attention models to detect counterfeit documents," in *Proc. Int. Conf. Document Anal. Recognit. (ICDAR)*, Sep. 2019, pp. 1332–1337.

[27] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.

[28] 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.

[29] R. Mudgalgundurao, P. Schuch, K. Raja, R. Ramachandra, and N. Damer, "Pixel-wise supervision for presentation attack detection on identity document cards," *IET Biometrics*, vol. 11, no. 5, pp. 383–395, Sep. 2022. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1049/bme2.12088

[30] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.

[31] C. Chen, S. Zhang, F. Lan, and J. Huang, "Domain-agnostic document authentication against practical recapturing attacks," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 2890–2905, 2022.

[32] Y. Viazovetskyi, V. Ivashkin, and E. Kashin, "StyleGAN2 distillation for feed-forward image manipulation," in *Computer Vision—ECCV*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham, Switzerland: Springer, 2020, pp. 170–186.

[33] J. Magee, S. Sheridan, and C. Thorpe, "An investigation into the application of the Meijering filter for document recapture detection," *J. Adv. Inf. Technol.*, vol. 15, no. 1, pp. 132–137, Jan. 2024. [Online]. Available: May 9, 2024. [Online]. Available: https://www.jait.us/show-235-1478-1.html, doi: 10.12720/jait.15.1.132-137.

[34] E. Meijering, M. Jacob, J. F. Sarria, P. Steiner, H. Hirling, and M. Unser, "Design and validation of a tool for neurite tracing and analysis in fluorescence microscopy images," *Cytometry A*, vol. 58A, no. 2, pp. 167–176, Apr. 2004.

[35] Á. D. S. Soares, R. B. Das Neves Junior, and B. L. D. Bezerra, "BID dataset: A challenge dataset for document processing tasks," in *Proc. Anais Estendidos da Conf. Graph., Patterns Images (SIBRAPI Estendido)*. Porto Alegre, Brasil: Sociedade Brasileira de Computação, Nov. 2020, pp. 143–146.

[36] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *Proc. Int. Conf. Comput. Vis.* Barcelona, Spain: IEEE, Nov. 2011, pp. 2564–2571. [Online]. Available: http://ieeexplore.ieee.org/document/6126544/

[37] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Computer Vision—ECCV 2016* (Lecture Notes in Computer Science), B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham, Switzerland: Springer, 2016, pp. 694–711. Accessed: May 9, 2024. [Online]. Available: http://link.springer.com/10.1007/978-3-319-46475-6_43

[38] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, Y. Bengio and Y. LeCun, Eds., May 2015, pp. 1–11.

[39] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[40] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–10. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7

[41] D. Schulz, J. Maureira, J. Tapia, and C. Busch, "Identity documents image quality assessment," in *Proc. 30th Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2022, pp. 1017–1021.

[42] R. Lara, A. Valenzuela, D. Schulz, J. Tapia, and C. Busch, "Towards an efficient semantic segmentation method of ID cards for verification systems," 2021, *arXiv:2111.12764*.

[43] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, and X. Chen, "Improved techniques for training GANs," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2016, pp. 2234–2242. [Online]. Available: https://papers.nips.cc/paper_files/paper/2016/file/8a3363abe792db2d8761d6403605aeb7-Metadata.json

[44] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2017, pp. 6626–6637. [Online]. Available: https://papers.nips.cc/paper_files/paper/2017/file/8a1d694707eb0fefe65871369074926d-Metadata.json

[45] M. Binkowski, D. J. Sutherland, M. Arbel, and A. Gretton, "Demystifying MMD GANs," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–15. [Online]. Available: https://openreview.net/forum?id=r1lUOzWCW

[46] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

[47] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *Proc. 5th Eur. Conf. Speech Commun. Technol. (Eurospeech)*. ISCA, Sep. 1997, pp. 1895–1898. Accessed: May 9, 2024. [Online]. Available: https://www.isca-archive.org/eurospeech_1997/martin97b_eurospeech.html

REUBEN P. MARKHAM received the B.Sc. degree in mathematics from the University of Alicante, in 2015, and the M.S. degree in statistical and computational information processing from the Complutense University of Madrid, in 2016. He is currently a Data Scientist with Instituto Tecnológico de Informática, where he has been involved in different client and research projects encompassing natural language processing, time series analysis, and computer vision. His main research interests include anomaly detection, probabilistic models, and applied research.

JUAN M. ESPÍN LÓPEZ received the B.Sc. degree in mathematics from the University of Murcia, in 2014, and the M.Sc. degree in applied mathematics, in 2015. He is currently pursuing the Ph.D. degree in computer science with the University of Murcia. He is a Senior Machine Learning Researcher with Facephi Biometria SA. His research interests include anti-spoofing systems for documents, face and voice, continuous authentication, speaker recognition, facial recognition, and machine learning and deep learning applications to the previous fields.

MARIO NIETO-HIDALGO received the B.Sc. degree in computer engineering, in 2011, the M.Sc. degree in technologies of information society, in 2012, and the Ph.D. degree in computer science from the University of Alicante, in 2017. He is currently a Senior Researcher with Facephi Biometria SA and an Adjunct Professor with the University of Alicante. His main research interests include computer vision and machine learning for face recognition and ID documents, presentation attack systems, gait analysis, and ambient assisted living.

JUAN E. TAPIA (Member, IEEE) received the P.E. degree in electronics engineering from Universidad Mayor, in 2004, and the M.S. and Ph.D. degrees in electrical engineering from the Department of Electrical Engineering, Universidad de Chile, in 2012 and 2016, respectively. In addition, he spent one year of internship with 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 with Universidad Andres Bello. From 2018 to 2020, he was the Research and Development Director for the electricity and electronics area with INACAP, Universidad Tecnologica de Chile, the Research and Development Director of TOC Biometrics Company, and an International Consultor on biometrics for face, iris applications and forensic/tampering ID-card detection. He is currently an Entrepreneur and a Senior Researcher with Hochschule Darmstadt (HDA), leading EU projects, such as iMARS and EINSTEIN. His main research interests include pattern recognition and deep learning applied to iris biometrics, morphing, feature fusion, and feature selection. He serves as a reviewer for a number of journals and conferences. He is on behalf of the German DIN Member of the ISO/IEC Sub-Committee 37 on biometrics.