GANprintR: Improved Fakes and Evaluation of the State of the Art in Face Manipulation Detection

João C. Neves, Ruben Tolosana, Ruben Vera-Rodriguez, Vasco Lopes, Hugo Proença, and Julian Fierrez

***Abstract*—The availability of large-scale facial databases, together with the remarkable progresses of deep learning technologies, in particular Generative Adversarial Networks (GANs), have led to the generation of extremely realistic fake facial content, raising obvious concerns about the potential for misuse. Such concerns have fostered the research on manipulation detection methods that, contrary to humans, have already achieved astonishing results in various scenarios. In this study, we focus on the synthesis of entire facial images, which is a specific type of facial manipulation. The main contributions of this study are four-fold: i) a novel strategy to remove GAN "fingerprints" from synthetic fake images based on autoencoders is described, in order to spoof facial manipulation detection systems while keeping the visual quality of the resulting images; ii) an in-depth analysis of the recent literature in facial manipulation detection; iii) a complete experimental assessment of this type of facial manipulation, considering the state-of-the-art fake detection systems (based on holistic deep networks, steganalysis, and local artifacts), remarking how challenging is this task in unconstrained scenarios; and finally iv) we announce a novel public database, named iFakeFaceDB, yielding from the application of our proposed GAN-fingerprint Removal approach (GANprintR) to already very realistic synthetic fake images. The results obtained in our empirical evaluation show that additional efforts are required to develop robust facial manipulation detection systems against unseen conditions and spoof techniques, such as the one proposed in this study.**

***Index Terms*—Fake news, face manipulation, face recognition, iFakeFaceDB, deepfakes, media forensics, GAN.**

I. Introduction

IMAGES and videos containing fake facial information obtained by digital manipulation have recently become a great public concern [1]. So far, the number and realism of digitally manipulated fake facial contents have been limited by the lack of sophisticated editing tools, the high domain of expertise required, and the complex and time-consuming process involved to generate realistic fakes. On the other hand, the scientific communities of biometrics and security in the past decade have been paying growing attention to understanding and protecting against what was considered a relevant threat around face biometrics [2]: presentation attacks conducted physically against the face sensor (camera) using various kinds of face spoofs (e.g., 2D or 3D printed, displayed, mask-based, etc.) [3], [4].

However, nowadays it is becoming increasingly easy to automatically synthesise non-existent faces or even to manipulate the face of a real person in an image/video, thanks to the free access to large public databases and also to the advances on deep learning techniques that eliminate the requirements of manual editing. As a result, accessible open software and mobile applications such as *ZAO* and *FaceApp* have led to large amounts of synthetically generated fake content [5], [6].

The most popular methods to generate fake face content can be categorised into four groups, regarding the level of manipulation [7]–[9]: *i)* entire face synthesis, *ii)* identity swap, *iii)* attribute manipulation, and *iv)* expression swap.

In this study, we focus on the entire face synthesis manipulation, where a machine learning model, typically based on Generative Adversarial Networks (GANs) [10], learns the distribution of the human face data, allowing to generate non-existent faces by sampling this distribution. This type of facial manipulation provides astonishing results, and is able to generate extremely realistic fakes. Nevertheless, contrary to humans, most state-of-the-art detection systems provide very good results against this type of facial manipulation, remarking how easy it is to detect the GAN "fingerprints" present in the synthetic images.

In this context, the main contributions of our paper are:

- A novel approach to spoof state-of-the-art facial manipulation detection systems, while keeping the visual quality of the resulting images. Fig. 1 graphically summarises our proposed approach based on a GAN-fingerprint Removal autoencoder (GANprintR).
- An in-depth literature analysis of the state-of-the-art detection approaches for the entire face synthesis manipulation, including the key aspects of the detection systems, the

Manuscript received November 12, 2019; revised April 3, 2020 and July 1, 2020; accepted July 1, 2020. Date of publication July 6, 2020; date of current version August 24, 2020. This work has been supported by projects: PRIMA (H2020-MSCA-ITN-2019-860315), TRESPASS-ETN (H2020-MSCA-ITN-2019-860813), BIBECA (RTI2018-101248-B-I00 MINECO/FEDER), BioGuard (Ayudas Fundación BBVA a Equipos de Investigación Cientíifica 2017), Accenture, by NOVA LINCS (UIDB/04516/2020) with the financial support of FCT - Fundação para a Ciência e a Tecnologia, through national funds, and by FCT/MCTES through national funds and co-funded by EU under the project UIDB/EEA/50008/2020. The guest editor coordinating the review of this manuscript and approving it for publication was Dr. Nasir Memon. *(João C. Neves and Ruben Tolosana contributed equally to this work.) (Corresponding author: Ruben Tolosana.)*

João C. Neves is with the NOVA-LINCS, 2829 Costa da Caparica, Portugal (e-mail: jcneves@di.ubi.pt).

Ruben Tolosana, Ruben Vera-Rodriguez, and Julian Fierrez are with the Biometrics and Data Pattern Analytics - BiDA Lab, Universidad Autonoma de Madrid, 28045 Madrid, Spain (e-mail: ruben.tolosana@uam.es; ruben.vera@uam.es; julian.fierrez@uam.es).

Vasco Lopes is with the University of Beira Interior, 6201-001 Covilhã, Portugal (e-mail: vasco.lopes@ubi.pt).

Hugo Proença is with the IT - Instituto de Telecomunicações, 3810-193 Aveiro, Portugal (e-mail: hugomcp@di.ubi.pt).

Digital Object Identifier 10.1109/JSTSP.2020.3007250

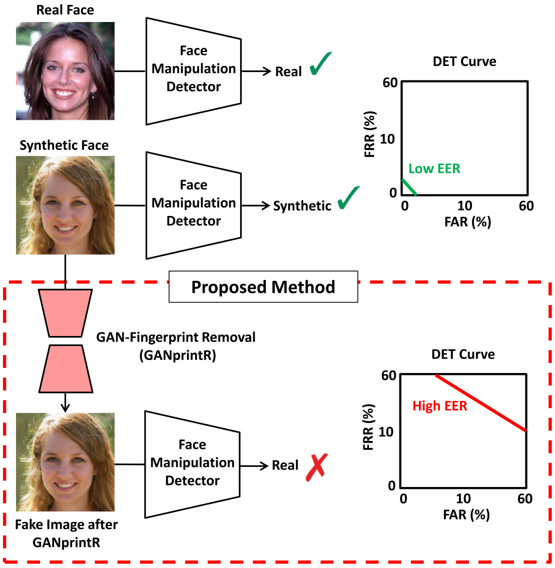


Fig. 1. Architecture of our proposed GAN-fingerprint removal approach. In general, state-of-the-art face manipulation detectors can easily distinguish between real and synthetic fake images. This usually happens due to the existence and exploitation by those detectors of GAN "fingerprints" produced during the generation of synthetic images. We propose an autoencoder module (GANprintR) to remove the GAN fingerprints from the synthetic images and spoof the facial manipulation detection systems, while keeping the visual quality of the resulting images.

databases used for developing and evaluating these systems, and the main results achieved by them.

- A thorough experimental assessment of this type of facial manipulation considering fake detection (based on holistic deep networks, steganalysis, and local artifacts) and realistic GAN-generated fakes (with and without the proposed GANprintR) over different experimental conditions, i.e., controlled and in-the-wild scenarios.
- We announce a novel database named iFakeFaceDB,[1] resulting from the application of our GANprintR to already very realistic synthetic images.

The remainder of the paper is organised as follows. Sec. II summarises previous studies focused on the detection of the entire face synthesis manipulation. Sec. III explains our proposed GAN-fingerprint removal approach. Sec. IV summarises the key features of the real and fake databases considered in our experimental framework. Sec. V and VI describe the proposed experimental setup and results achieved, respectively. Finally, Sec. VII draws the final conclusions and points out some lines for future work.

II. Related Work

Various studies have recently evaluated how easy it is to detect manipulations based on the entire face synthesis. Table I shows a comparison of the most relevant approaches in this area. For each study, we include information related to the features, classifiers, best performance, and databases considered.

In [10], the authors analysed the architecture of GANs in order to detect different artifacts between fake and real images. They proposed a detection system based on colour features and a linear Support Vector Machine (SVM) for the final classification. Their approach achieved a final 70% Area Under the Curve (AUC) for the best performance when considering the NIST MFC2018 dataset [19]. A similar approach was followed by Matern *et al.* [15] where the authors exploited relatively simple visual artifacts from specific facial regions (e.g., eyes, teeth, facial contours) to detect different types of facial manipulations. In a similar research line, Yang *et al.* [14] exploited the weakness of GANs in generating consistent head poses, and trained a SVM to distinguish between real and synthetic faces based on the estimation of the 3D head pose.

In [16], the authors exploited different color channels (YCbCr, HSV and Lab) to extract from a Convolutional Neural Network (CNN) different deep representations, which were subsequently fed to a Random Forest classifier for deciding the realness of an image.

In [13], Wang *et al.* conjectured that monitoring neuron behavior could also serve as an asset in detecting fake faces since layer-by-layer neuron activation patterns may capture more subtle features that are important for the facial manipulation detection system. Their proposed approach, named FakeSpoter, extracted as features neuron coverage behaviors of real and fake faces from deep face recognition systems (i.e., VGG-Face [20], OpenFace [21], and FaceNet [22]), and then trained a SVM for the final classification. The authors tested their proposed approach using real faces from CelebA-HQ [23] and FFHQ [24] databases and synthetic faces created through InterFaceGAN [25] and StyleGAN [24], achieving for the best performance a final 84.78% accuracy for the FaceNet model.

More recently, Stehouwer *et al.* carried out in [9] a complete analysis of different facial manipulation detection methods. They proposed to use attention mechanisms to process and improve the feature maps of CNN models. For the facial manipulation method considered in our study (i.e., entire face synthesis), the authors achieved a final 0.05% Equal Error Rate (EER) considering real faces from CelebA [26], FFHQ [24], and FaceForensics++ [27] databases and fake images created through PGGAN [28] and StyleGAN [24] approaches.

Wang *et al.* carried out in [17] a very interesting research using publicly available commercial software from Adobe Photoshop in order to synthesise new faces [29], and also a professional artist in order to manipulate 50 real photographs. The authors began running a human study through Amazon Mechanical Turk (AMT), showing real and fake images to the participants and asking them to classify each image into one of the classes. The results remark the task difficulty for humans, with a final 53.5% performance (chance = 50%). After the human study, the authors proposed an automatic detection system based on Dilated Residual Networks (DRN), achieving Average Precisions (AP) of 99.8% and 97.4% for automatic and manual face synthesis manipulation detection.

In another line of research, some authors have recently focused on the problem of finding the GAN architecture used for generating a specific image potentially synthetic [30], [31].

[1] [Online]. Available: https://github.com/socialabubi/iFakeFaceDB

TABLE I
COMPARISON OF STATE-OF-THE-ART MANIPULATION DETECTION APPROACHES FOR ENTIRE FACE SYNTHESIS MANIPULATION

Study	Features	Classifiers	Best Performance	Databases
McCloskey and Albright (2018) [10]	Color-related Features	SVM	AUC = 70%	NIST MFC2018
Yu *et al.* (2018) [11]	GAN-related Features	CNN	Acc. = 99.50%	Real: CelebA Fake: Own Database
Marra *et al.* (2018) [12]	Image-related Features	CNN	Acc. = 95.07%	Real: Own Database(CycleGAN) Fake: Own Database(CycleGAN)
Wang *et al.* (2019) [13]	CNN Neuron Behavior Features	SVM	Acc. = 84.78%	Real: CelebA-HQ/FFHQ Fake: Own Database
Stehouwer *et al.* (2019) [9]	Image-related Features	CNN + Attention Mechanism	EER = 0.05%	Real: CelebA/FFHQ/FaceForensics++ Fake: Own Database
Yang *et al.* (2019) [14]	Head Pose	SVM	AUC = 89%	Real: UADFV/DARPA MediFor Fake: UADFV/DARPA MediFor
Matern *et al.* (2019) [15]	Eye Color Features	K-NN	AUC = 85.2%	Real: CelebA Fake: Own Database (PGGAN)
He *et al.* (2019) [16]	Color-related Features	Random Forest	Acc. = 99%	Real: CelebA Fake: Own Database (PGGAN)
Wang *et al.* (2019) [17]	Image-related Features	DRN	AP = 99.8%	Real: Own Database Fake: Own Database

Yu *et al.* analysed in [11] the existence and uniqueness of GAN fingerprints to detect fake images. In particular, they proposed a learning-based formulation consisting of an attribution network architecture to map an input image to its corresponding fingerprint image. Therefore, they learned a model fingerprint for each source (each GAN instance plus the real world), such that the correlation index between one image fingerprint and each model fingerprint serves as softmax logit for classification. Their proposed approach was tested using real faces from CelebA database [26] and synthetic faces created through different GAN approaches (PGGAN [28], SNGAN [32], CramerGAN [33], and MMDGAN [34]), achieving a final 99.50% accuracy for the best performance in manipulation detection.

Finally, we also include for completeness some relevant references to other recent studies focused on the detection of general GAN-based image manipulations, not facial ones: [12], [35]–[38].

III. PROPOSED APPROACH: GAN-FINGERPRINT REMOVAL (GANPRINTR)

Our approach aims at transforming synthetic face images, such that their visual appearance is unaltered but the GAN fingerprints (the discriminative information that permits the distinction from real imagery) are removed. Considering that the fingerprints are high frequency signals [31], we hypothesise that their removal could be performed by an autoencoder, which acts as a non-linear low-pass filter. We claim that by using this strategy, the detection capability of state-of-the-art facial manipulation detection methods significantly decreases, while at the same time humans still are not capable of perceiving that images were transformed.

In general, an autoenconder comprises two distinct networks, encoder ψ and decoder γ:

$$\begin{aligned} \psi &: X \mapsto l \\ \gamma &: l \mapsto X' \end{aligned} \tag{1}$$

where X denotes the input image to the network, l is the latent feature representation of the input image after passing through the encoder ψ, and X' is the reconstructed image generated from l, after passing through the decoder γ. The networks ψ and γ can be learned by minimising the reconstruction loss $L_{\psi,\gamma}(X, X') = ||X - X'||^2$ over a development dataset following an iterative learning strategy.

As result, when L is nearly 0, ψ is able to discard all redundant information from X and code it properly into l. However, for a reduced size of the latent feature representation vector, L will increase and ψ will be forced to encode in l only the most representative information of X. We claim this kind of autoencoder acts as a GAN-fingerprint removal system.

Fig. 2 describes our proposed approach based on a convolutional AutoEncoder (AE) composed of a sequence of 3 × 3 convolutional filters, coupled with ReLU activation functions. After each convolutional layer, a 2 × 2 max-pooling layer is used to progressively decrease the size of the activation map to $28\times28\times8$, which represents the bottleneck of the reconstruction model.

The AE is trained with images from a public dataset that comprises face imagery from real persons. In the evaluation phase, the AE is used to generate improved fakes from input fake faces where GAN "fingerprints", if present in the initial fakes, will be reduced. The main rationale of this strategy is that by training with real images the AE can learn the core structure of this type of natural data, which can then be exploited to improve existing fakes.

IV. DATABASES

Four different public databases and one generated are considered in the experimental framework. Fig. 3 shows some examples of each database. We now summarise the most important features.

A. Real Face Images

1) CASIA-WebFace [39]: this database contains 494,414 face images from 10,575 actors and actresses of IMDb. Face images comprise random pose variations, illumination, facial expression, and resolution.

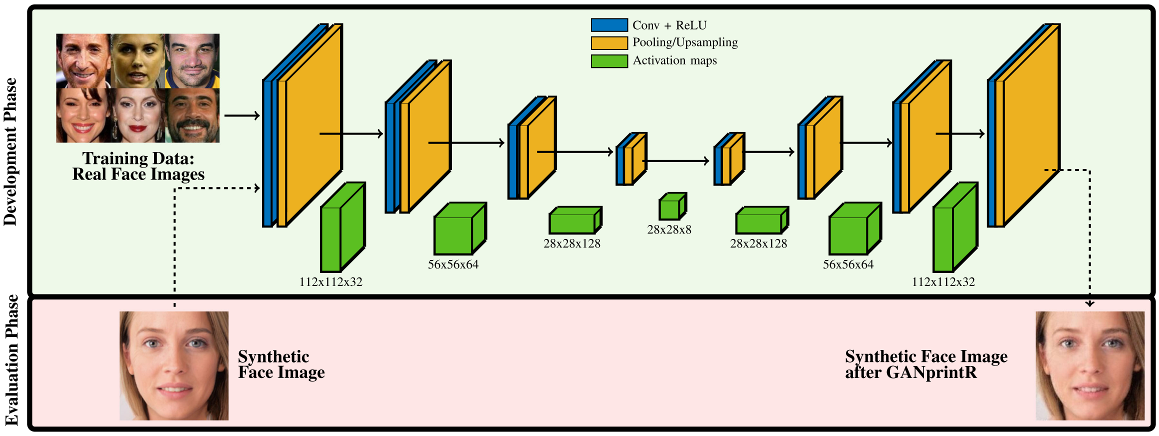


Fig. 2. Proposed GAN-fingerprint Removal module (GANprintR) based on a convolutional AutoEncoder (AE). The AE is trained using only real face images from the development dataset. In the evaluation stage, once the autoencoder is trained, we can pass synthetic face images through it to provide them with additional naturalness, in this way removing the GAN-fingerprint information that may be present in the initial fakes.

2) VGGFace2 [40]: this database contains 3.31 million images from 9,131 different subjects, with an average of 363 images per subject. Images were downloaded from the Internet and contain large variations in pose, age, illumination, ethnicity, and profession (e.g., actors, athletes, and politicians).

B. Synthetic Face Images

1) TPDNE: this database comprises 150,000 unique faces, collected from the website.[2] Synthetic images are based on the recent StyleGAN approach [24] trained with FFHQ database [41].

2) 100K-Faces [42]: this database contains 100,000 synthetic images generated using StyleGAN [24]. In this database the StyleGAN network was trained using around 29,000 photos of 69 different models, producing face images with a flat background.

3) PGGAN [28]: this database comprises 80,000 synthetic face images generated using the PGGAN network. In particular, we consider the publicly available model trained using the CelebA-HQ database.

V. Experimental Setup

A. Pre-Processing

In order to ensure fairness in our experimental validation, we created a curated version of all the datasets where the confounding variables were removed. Two different factors were considered in this study:

- *Background*: this is a clearly distinctive aspect among real and synthetic face images as different acquisition conditions are considered in each database.
- *Head pose*: images generated by GANs hardly ever produce high variation from the frontal pose [9], contrasting with most popular real face databases such as CASIA-WebFace and VGGFace2. Therefore, this factor may falsely improve the performance of the detection systems since non-frontal images are more likely to be real faces.

To remove these factors from both the real and synthetic images, we extracted 68 face landmarks, using the method described in [43]. Given the landmarks of the eyes, an affine transformation was determined such that the location of the eyes appears in all images at the same distance from the borders. This step allowed to remove all the background information of the images while keeping the maximum amount of the facial regions. Regarding the head pose, landmarks were used to estimate the pose (*frontal* vs. *non-frontal*). In our experimental framework, we kept only the frontal face images, in order to avoid biased results. After this pre-processing stage, we were able to provide images of constant size (224 × 224 pixels) as input to the systems. Fig. 3 shows examples of the crop-out faces of each database after applying the pre-processing steps. The synthetic images obtained by this pre-processing stage are the ones used to create the database iFakeFaceDB after being processed by our GANprintR approach.

B. Facial Manipulation Detection Systems

Three different state-of-the-art manipulation detection approaches are considered in this study.

1) *XceptionNet* [44]: this network was selected, essentially because it provides the best detection results in the most recently published studies [9], [27], [45]. We followed the same training approach considered in [27]: *i)* the model was initialized with the weights obtained after training with the ImageNet dataset [46], *ii)* we changed the last fully-connected layer of the ImageNet model by a new one (two classes, real or synthetic image), *iii)* we fixed all weights up to the final layers and pre-trained the network for few epochs, and finally *iv)* we trained the network for 20 more epochs and chose the best performing model based on validation accuracy.

2) *Steganalysis* [38]: the method by Nataraj *et al.* was selected for providing an approach based on steganalysis, rather than directly extracting features from the images, as in the

[2] [Online]. Available: https://thispersondoesnotexist.com

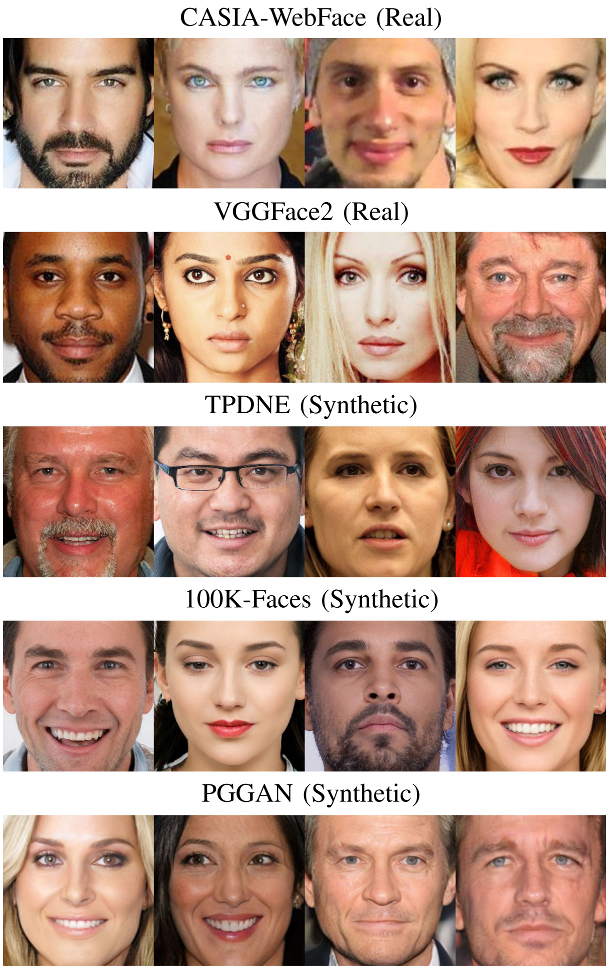
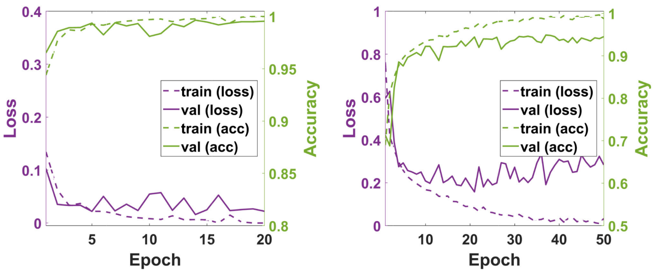


Fig. 3. Examples of the databases considered in our experiments after applying the pre-processing stage described in Sec. V-A.

XceptionNet approach. In particular, this approach calculates the co-occurrence matrices directly from the image pixels on each channel (red, green and blue), and passes this information through a custom CNN, which allows the network to extract non-linear robust features. Considering that the source code is not available from the authors, we replicated this technique to perform our experiments.

3) *Local Artifacts* [15]: we have chosen the method of Matern *et al.*, because it provides an approach based on the direct analysis of the visual facial artifacts, in opposition to the remaining approaches that follow holistic strategies. In particular, the authors of that work claim that some parts of the face (e.g., eyes, teeth, facial contours) provide useful information about the authenticity of the image, and thus train a classifier to distinguish between real and synthetic face images using features extracted from these facial regions.

All our experiments were implemented under a PyTorch framework, with a NVIDIA Titan X GPU. The training of the Xception network was performed using the Adam optimiser with a learning rate of 10^{-3}, dropout for model regularization with a rate of 0.5, and a binary cross-entropy loss function. Regarding the steganalysis approach, we reused the parameters adopted for Xception network, since the authors of [38] did not detail the training strategy adopted. Regarding the local artifacts approach, we adopted the strategy for detecting "generated faces", where a k-nearest neighbour classifier is used to distinguish between real and synthetic face images based on eye color features.



(a) XceptionNet [44] (b) Steganalysis [38]

Fig. 4. **Exp. A.1:** evolution of the loss/accuracy with the number of epochs.

C. Protocol

The experimental protocol designed in this study aims at performing an exhaustive analysis of state-of-the-art facial manipulation detection systems. As such, three different experiments are considered: *i)* controlled scenarios, *ii)* in-the-wild scenarios, and *iii)* GAN-fingerprint removal.

Each database was divided into two disjoint datasets, one for the development of the systems (70%) and the other one for evaluation purposes (30%). Additionally, the development dataset was divided into two disjoint subsets, training (75%) and validation (25%). The same number of real and synthetic images were considered in the experimental framework. In addition, for real face images, different users were considered in the development and evaluation datasets, in order to avoid biased results.

Our proposed GANprintR was trained during 100 epochs, using the Adam optimizer with a learning rate of 10^{-3}, and mean square error (MSE) to obtain the reconstruction loss. To ensure an unbiased evaluation, our GANprintR was trained with images from the MS-Celeb dataset [47], since it is disjoint from the datasets used in the development and evaluation of all the fake detection systems used in our experiments.

VI. Experimental Results

A. Controlled Scenarios

In this section, we report the results of the detection of entire face synthesis in controlled scenarios, i.e., when samples from the same databases were considered for both development and final evaluation of the detection systems. This is the strategy commonly used in most studies, typically resulting in very good performance (see Sec. II).

A total of six experiments are carried out: A.1 to A.6. Table II describes the development and evaluation databases considered in each experiment together with the corresponding final evaluation results in terms of EER. Additionally, we represent in Fig. 4 the evolution of the loss/accuracy of the XceptionNet and Steganalysis detection systems for Exp. A.1.

TABLE II

CONTROLLED AND IN-THE-WILD SCENARIOS: MANIPULATION DETECTION PERFORMANCE IN TERMS OF EER AND RECALL (%) FOR DIFFERENT DEVELOPMENT AND EVALUATION SETUPS. R_{real} AND R_{fake} DENOTE THE RECALL OF THE REAL AND FAKE CLASSES, RESPECTIVELY. CONTROLLED (EXP. A.1-A.6). IN-THE-WILD (EXP. B.1-B.24). VF2 = VGGFACE2. CASIA = CASIA-WEBFACE. ALL METRICS ARE GIVEN IN (%)

	Development		**Evaluation**		**XceptionNet** [44]			**Steganalysis** [38]			**Local Artifacts** [15]		
Experiment	Real	Synthetic	Real	Synthetic	EER	R_{real}	R_{fake}	EER	R_{real}	R_{fake}	EER	R_{real}	R_{fake}
A.1	VF2	TPDNE	VF2	TPDNE	0.22	99.77	99.80	10.92	89.07	89.10	38.53	60.72	62.20
B.1	VF2	TPDNE	VF2	100F	0.45	99.30	99.80	23.07	71.66	85.59	35.86	64.13	64.16
B.2	VF2	TPDNE	VF2	PGGAN	13.82	78.44	99.73	27.12	67.28	83.87	40.10	59.05	60.80
B.3	VF2	TPDNE	CASIA	100F	0.35	99.30	100.00	24.00	71.23	83.53	35.61	64.05	64.69
B.4	VF2	TPDNE	CASIA	PGGAN	13.72	78.47	100.00	28.05	66.81	81.61	39.87	59.0	61.4
A.2	VF2	100F	VF2	100F	0.28	99.70	99.73	12.28	87.70	87.73	31.45	67.83	69.26
B.5	VF2	100F	VF2	TPDNE	21.18	70.32	99.54	28.02	66.72	82.09	42.89	55.17	60.16
B.6	VF2	100F	VF2	PGGAN	44.43	52.96	97.71	32.62	62.35	79.31	48.70	50.53	52.87
B.7	VF2	100F	CASIA	TPDNE	21.07	70.37	99.94	28.85	66.29	80.14	46.04	52.50	55.98
B.8	VF2	100F	CASIA	PGGAN	44.32	53.01	99.71	33.45	61.90	77.15	51.89	47.8	48.6
A.3	VF2	PGGAN	VF2	PGGAN	0.02	99.97	100.00	3.32	96.67	96.70	35.13	64.33	65.41
B.9	VF2	PGGAN	VF2	TPDNE	16.85	74.79	100.00	33.32	60.42	91.74	40.84	57.55	61.17
B.10	VF2	PGGAN	VF2	100F	5.85	89.53	100.00	25.60	66.87	94.04	44.47	53.99	57.77
B.11	VF2	PGGAN	CASIA	TPDNE	16.85	74.79	100.00	35.73	59.19	81.85	39.89	58.02	62.82
B.12	VF2	PGGAN	CASIA	100F	5.85	89.53	100.00	28.02	65.73	86.50	43.53	54.5	59.5
A.4	CASIA	TPDNE	CASIA	TPDNE	0.02	99.97	100.00	12.08	87.90	87.93	39.36	59.62	61.65
B.13	CASIA	TPDNE	VF2	100F	1.75	99.35	97.20	36.68	59.58	71.82	39.03	60.67	61.25
B.14	CASIA	TPDNE	VF2	PGGAN	4.42	94.21	97.04	30.77	65.13	76.40	38.94	61.02	61.10
B.15	CASIA	TPDNE	CASIA	100F	0.32	99.37	100.00	34.12	61.02	78.41	38.05	61.20	62.67
B.16	CASIA	TPDNE	CASIA	PGGAN	2.98	94.37	100.00	28.20	66.48	82.19	37.96	61.5	62.5
A.5	CASIA	100F	CASIA	100F	0.08	99.90	99.93	16.05	83.94	83.96	33.96	65.04	67.03
B.17	CASIA	100F	VF2	TPDNE	5.93	97.69	90.95	34.00	62.64	71.80	43.11	55.00	59.83
B.18	CASIA	100F	VF2	PGGAN	10.08	89.64	90.20	45.63	52.91	58.71	46.36	52.37	55.92
B.19	CASIA	100F	CASIA	TPDNE	1.10	97.91	99.93	31.67	63.97	76.67	44.22	53.94	58.54
B.20	CASIA	100F	CASIA	PGGAN	5.25	90.55	99.93	43.30	54.34	64.74	47.49	51.3	54.6
A.6	CASIA	PGGAN	CASIA	PGGAN	0.05	99.93	99.97	4.62	95.37	95.40	34.79	64.42	66.00
B.21	CASIA	PGGAN	VF2	TPDNE	4.90	99.96	91.10	31.73	61.93	88.92	43.52	55.25	57.94
B.22	CASIA	PGGAN	VF2	100F	4.88	100.00	91.10	41.97	54.63	80.35	44.69	54.05	56.89
B.23	CASIA	PGGAN	CASIA	TPDNE	0.03	99.97	99.97	31.43	62.08	90.07	41.46	56.64	61.00
B.24	CASIA	PGGAN	CASIA	100F	0.02	100.00	99.97	41.67	54.79	82.22	42.63	55.5	60.0

The analysis of Fig. 4 shows that both XceptionNet and Steganalysis approaches are able to learn discriminative features to detect between real and synthetic face images. The training process was faster for the XceptionNet detection system compared with Steganalysis, converging to a lower loss value in fewer epochs (close to zero after 20 epochs). The best validation accuracies achieved in Exp. A.1 for the XceptionNet and Steganalysis approaches are 99% and 95%, respectively. Similar trends are observed for the other experiments.

We now analyse the results included in Table II for experiments A.1 to A.6. Analysing the results obtained by the XceptionNet system, almost ideal performance is achieved with EER values less than 0.5%. These results are in agreement to previous studies in the topic (see Sec. II), pointing for the potential of the XceptionNet model in controlled scenarios. Regarding the Steganalysis approach, a higher degradation of the system performance is observed, when compared with the XceptionNet approach, especially for the 100K-Face database, e.g., a 16% EER is obtained in Exp. A.5. Finally, it can be observed that the approach based on local artifacts was the least efficient to spot the differences between real and synthetic data, with an average 35.5% EER over all experiments.

In summary, for controlled scenarios XceptionNet has excellent manipulation detection accuracies, then Steganalysis provides good accuracies, and finally Local Artifacts has poor accuracy. In the next section we will see the limitations of these techniques in-the-wild.

B. In-the-Wild Scenarios

This section evaluates the performance of the facial manipulation detection systems in more realistic scenarios, i.e., in-the-wild. The following aspects are considered: *i)* different development and evaluation databases, and *ii)* different image resolution/blur among the development and evaluation of the models. This last point is particularly important, as the quality of raw images/videos is usually modified when, e.g., they are uploaded to social media. The effect of image resolution has been preliminary analysed in previous studies [27], [48], but for different facial manipulation groups, i.e., face swapping/identity swap and facial expression manipulation. The main goal of this section is to analyse the generalisation capability of state-of-the-art entire face synthesis detection in unconstrained scenarios.

First, we focus on the scenario of considering the same real but different synthetic databases in development and evaluation (Exp. B.1, B.2, B.5, B.6, and so on, provided in Table II). In general, the results achieved in the experiments evidence a high degradation of the detection performance regardless of the

TABLE III

Comparison Between the Proposed Approach (GANprintR) and Typical Image Manipulations. The Detection Performance is Provided in Terms of EER and Recall (%) for Experiments A.1 to A.6, When Using Different Versions of the Evaluation Set. TED Stands for Transformation of the Evaluation Data and Details the Technique Used to Modify the Test Set Before Fake Detection. R_{real} and R_{fake} Denote the Recall of the Real and Fake Classes, Respectively

	Development		**Evaluation**		**XceptionNet** [44]					
Experiment	Real	Synthetic	Real	Synthetic	TED	EER(%)	R_{real}(%)	R_{fake}(%)	PSNR(db)	SSIM
A.1	VF2	TPDNE	VF2	TPDNE	Original	0.22	99.77	99.80	-	-
					Downsize	1.17	98.83	98.87	35.55	0.93
					Low-Pass Filter	0.83	99.17	99.20	34.63	0.92
					JPEG Compression	1.53	98.47	98.50	36.02	0.96
					GANprintR	10.63	89.37	89.40	35.01	0.96
A.2	VF2	100F	VF2	100F	Original	0.28	99.70	99.73	-	-
					Downsize	0.87	99.13	99.17	36.24	0.95
					Low-Pass Filter	2.87	97.10	97.13	35.22	0.93
					JPEG Compression	1.83	98.17	98.20	36.76	0.97
					GANprintR	6.37	93.64	93.66	35.59	0.96
A.3	VF2	PGGAN	VF2	PGGAN	Original	0.02	99.97	100.00	-	-
					Downsize	3.70	96.27	96.30	34.85	0.91
					Low-Pass Filter	1.53	98.43	98.47	34.10	0.90
					JPEG Compression	30.93	69.04	69.06	35.85	0.96
					GANprintR	17.27	82.71	82.73	34.82	0.95
A.4	CASIA	TPDNE	CASIA	TPDNE	Original	0.02	99.97	100.00	-	-
					Downsize	1.00	98.97	99.00	35.55	0.93
					Low-Pass Filter	0.07	99.90	99.93	34.63	0.92
					JPEG Compression	2.50	97.47	97.50	36.02	0.96
					GANprintR	4.47	95.50	95.53	35.01	0.96
A.5	CASIA	100F	CASIA	100F	Original	0.08	99.90	99.93	-	-
					Downsize	6.27	93.70	93.73	36.24	0.95
					Low-Pass Filter	11.53	88.44	88.46	35.22	0.93
					JPEG Compression	3.27	96.73	96.77	36.76	0.97
					GANprintR	11.47	88.50	88.53	35.59	0.96
A.6	CASIA	PGGAN	CASIA	PGGAN	Original	0.05	99.93	99.97	-	-
					Downsize	7.77	92.24	92.26	34.85	0.91
					Low-Pass Filter	2.10	97.90	97.93	34.10	0.90
					JPEG Compression	5.37	94.64	94.66	35.85	0.96
					GANprintR	8.37	91.64	91.66	34.82	0.95

facial manipulation detection approach. For the XceptionNet, the average EER is 11.2%, i.e., over 20 times higher than the results achieved in Exp. A.1-A.6 (<0.5% average EER). Regarding the Steganalysis approach, the average EER is 32.5%, i.e., more than 3 times higher than the results achieved in Exp. A.1-A.6 (9.8% average EER). For Local Artifacts, the observed average EER is 42.4%, with an average worsening of 19%. The large degradation of the first two detectors suggests that they might rely heavily on the GAN fingerprints of the training data. This result confirms the hypothesis that different GAN models produce different fingerprints, as also mentioned in previous studies [11]. Moreover, these results suggest that these GAN fingerprints are the information used by the detectors to distinguish between real and synthetic data.

Table II also considers the case of using different real and synthetic databases for both development and evaluation (Exp. B.3, B.4, B.7, B.8, etc.). In this scenario, average EERs of 9.3%, 32.3% and 42.3% in fake detection are obtained for XceptionNet, Steganalysis, and Local Artifacts, respectively. When comparing these results with the EERs of the previous experiments (where only the synthetic evaluation set was changed), no significant gap in performance is found, which suggests that the change of synthetic data in training might be the main cause for performance degradation.

Finally, we also analyse how different image transformations affect facial manipulation detection systems. In this analysis, we focus only on the XceptionNet model as it provides much better results when compared with the remaining detection systems. For each baseline experiment (A.1 to A.6), the evaluation set (both real and fake images) was transformed by: *i)* resolution downsizing (1/3 of the original resolution), *ii)* a low-pass filter (9×9 Gaussian kernel, $\sigma = 1.7$), and *iii)* jpeg image compression using a quality level of 60. The resulting EER together with the Recall, PSRN, and SSIM values are provided in Table III, together with the performance of the original images. The results suggest a high performance degradation in manipulation detection for all experiments, proving the vulnerability of fake detection systems to unseen conditions, even if they result from simple image transformations. These findings agree with the conclusions extracted in other studies of the literature. For example, Marra *et al.* evaluated in [12] the robustness of different fake detectors over different training and testing scenarios, considering fake images created using image-to-image translations [49]. For the XceptionNet approach and the image compression scenario, the authors achieved an accuracy in manipulation detection of 87.17%, an average absolute worsening of 7.32% compared with the uncompressed scenario (accuracy of 94.49%).

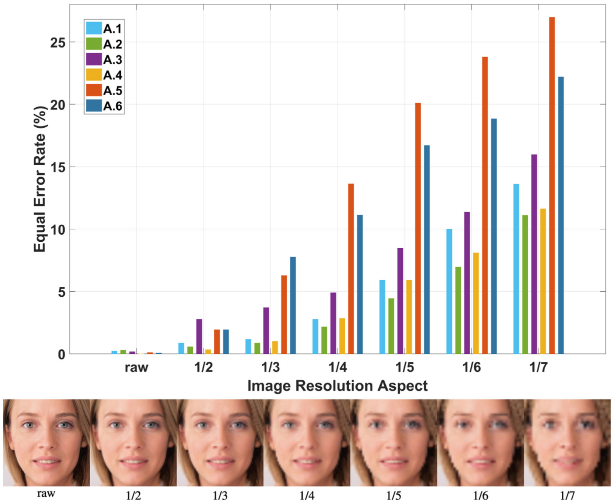


Fig. 5. Robustness of the fake detection system regarding the image resolution. The XceptionNet model is trained with the raw image resolution and evaluated with lower image resolutions. Note how the EER increases significantly while reducing the image resolution.

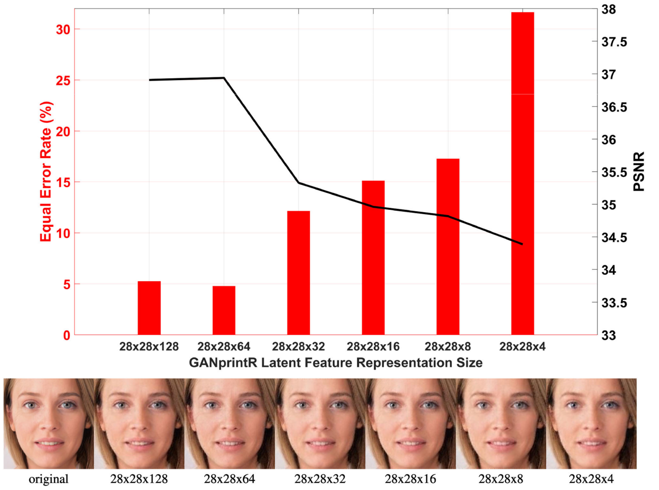


Fig. 6. Robustness of the fake detection system after our proposed GAN-fingerprint Removal (GANprintR). The latent feature representation size of the AE is varied to analyse the impact on both system performance and visual aspect of the reconstructed images. Note how the EER increases significantly when considering our proposed spoof approach, while maintaining a high visual similarity with the original image.

To further understand the impact of these transformations, we evaluated an increasing downsize ratio in the performance of the fake detection system. Fig. 5 depicts the detection performance results in terms of EER(%), from lower to higher modifications of the image resolution. In general, we can observe increasingly higher degradation of the fake detection performance for decreasing resolution. For example, when the image resolution is reduced to 1/4, the average EER in manipulation detection increases 6% when compared with the raw image resolution (raw equals to 1/1). This performance degradation is even higher when we further reduce the image resolution, with EERs(%) higher than 15%. These results support the conclusion about a poor generalisation capacity of state-of-the-art facial manipulation detection systems to unseen conditions.

C. GAN-Fingerprint Removal

This section analyses the results of the proposed strategy for GAN-fingerprint Removal (GANprintR). We evaluated to what extent our method is capable of spoofing state-of-the-art fake detectors by improving fake images already obtained with some of the best and most realistic known methods for entire face synthesis. For this, the experiments A.1 to A.6 were repeated for the XceptionNet detection system, but the fake images of the evaluation set were transformed after passing through our proposed GANprintR. Table III provides the results achieved for both the original fake data and after GANprintR. The analysis shows that GANprintR results in higher fake detection error than the remaining attacks, while maintaining a similar or even better visual quality. In all the experiments, the EER of the manipulation detection increases when using GANprintR to transform the synthetic face images. Also, the detection degradation is higher than other types of attacks for similar PSNR values and slightly higher values of SSIM. In particular, the average EER when considering GANprintR is 9.8%, i.e., over 20 times higher than the results achieved when using the original fakes (<0.5% average EER). This suggests that our method is not simply removing high-frequency information (evidenced by the comparison with the low-pass filter and downsize) but it is also removing the GAN fingerprints from the fakes improving their naturalness. It is important to remark that different real face databases were considered for training the face manipulation detection systems and our GANprintR module.

In addition, we provide in Fig. 6 an analysis of the impact of the latent feature representation of the autoencoder in terms of EER and PSNR. In particular, we follow the experimental protocol considered in Exp. A.3, and calculate the EER of XceptionNet for detecting fakes improved with various configurations of GANprintR. Moreover, the PSNR for each set of transformed images is also included in Fig. 6 together with a face example of each configuration to visualise the image quality. The face examples included in Fig. 6 show no substantial differences between the original fake and the resulting fakes after GANprintR for the different latent feature representation size of the GANprintR, which is confirmed by the tight range of PSNR values obtained along the different latent feature representations. On the other hand, EER values of fake detection significantly increase as the size of latent feature representations diminish, evidencing that GANprintR is capable of spoofing state-of-the-art manipulation detection systems without significantly degrading the visual aspect of the image.

Finally, to confirm that GANprintR is actually removing the GAN-fingerprint information and not just reducing the image resolution of the images, we performed a final experiment where we trained the XceptionNet for fake detection considering different levels of image resolution, and then tested it using fakes improved with GANprintR. Fig. 7 shows the fake detection performance in terms of EER for different size of the latent feature representation of GANprintR. Five different GANprintR

TABLE IV

Impact of the GANprintR Approach on Three State-of-the-Art Manipulation Detection Approaches. A Significant Performance Degradation is Observed in All Manipulation Detection Approaches When Exposed to Images Transformed by the Proposed GANprintR Approach. The Detection Performance is Provided in Terms of EER and Recall (%), While R_{real} and R_{fake} Denote the Recall of the Real and Fake Classes, Respectively

	Development		Evaluation		XceptionNet [44]				Steganalysis [38]			Local Artifacts [15]		
Experiment	Real	Synthetic	Real	Synthetic	data	EER(%)	R_{real}(%)	R_{fake}(%)	EER(%)	R_{real}(%)	R_{fake}(%)	EER(%)	R_{real}(%)	R_{fake}(%)
A.1	VF2	TPDNE	VF2	TPDNE	Original	0.22	99.77	99.80	10.92	89.07	89.10	38.53	60.72	62.20
					GANprintR	10.63	89.37	89.40	22.37	77.61	77.63	44.06	55.16	56.67
A.2	VF2	100F	VF2	100F	Original	0.28	99.70	99.73	12.28	87.70	87.73	31.45	67.83	69.26
					GANprintR	6.37	93.64	93.66	17.30	82.71	82.73	36.35	62.93	64.41
A.3	VF2	PGGAN	VF2	PGGAN	Original	0.02	99.97	100.00	3.32	96.67	96.70	35.13	64.33	65.41
					GANprintR	17.27	82.71	82.73	35.13	64.85	64.85	42.24	57.28	58.29
A.4	CASIA	TPDNE	CASIA	TPDNE	Original	0.02	99.97	100.00	12.08	87.90	87.93	39.36	59.62	61.65
					GANprintR	4.47	95.50	95.53	24.97	75.04	75.06	42.75	56.16	58.37
A.5	CASIA	100F	CASIA	100F	Original	0.08	99.90	99.93	16.05	83.94	83.96	33.96	65.04	67.03
					GANprintR	11.47	98.50	98.53	19.80	80.17	80.19	38.14	60.77	62.97
A.6	CASIA	PGGAN	CASIA	PGGAN	Original	0.05	99.93	99.97	4.62	95.37	95.40	34.79	64.42	66.00
					GANprintR	8.37	93.64	93.66	27.77	72.21	72.22	39.15	60.02	61.70

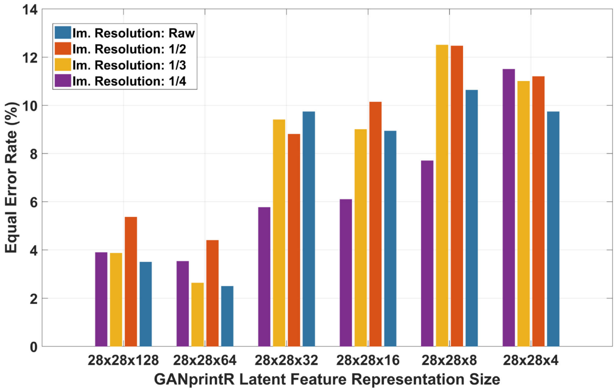


Fig. 7. Robustness of the fake detection system trained with different resolutions and then tested with fakes improved with GANprintR under various configurations (representation sizes). Five different GANprintR configurations are tested per image resolution level. The results observed point for the stability of EER values with respect to using downsized synthetic images in training. This observation supports the conclusion that GANprintR is actually removing the GAN-fingerprint information.

configurations are tested per image resolution. The obtained results point for the stability of EER values with respect to downsized synthetic images in training, concluding that our proposed approach is actually removing the GAN-fingerprint information.

D. Impact of GANprintR on Other Fake Detectors

For completeness, this section provides a comparative analysis of the impact of GANprintR on the three state-of-the-art fake detectors considered in this study. Table IV reports the EER and Recall observed when using the original fake images and the same ones after passing through GANprintR.

In general, the same conclusions highlighted for XceptionNet in Sec. VI-C are extracted in Table IV for the other two fake detectors. For XceptionNet, an average absolute worsening of 9.65% EER is produced when using GANprintR. This degradation is even higher for the Steganalysis fake detector with an average absolute worsening of 14.68% EER. Finally, the fake detector based on Local Artifacts has proven to be the most robust one, with an average absolute worsening of 4.91% EER. This lower performance degradation can be produced due to the higher EERs achieved in the original fake images (an average 35.54% EER). These performance degradations prove the success of our proposed GANprintR, creating improved versions of the original fake images.

VII. Conclusion

In this paper we presented a method (GANprintR) for improving the naturalness of facial fake images based on autoencoders, and we have empirically shown its ability to deceive state-of-the-art manipulation detection methods in a larger extent than some of the most sophisticate and realistic GAN-based synthetic face image generators available in the literature. Our method and experiments have been positioned and discussed in comparison with key related works around this problem published in the last couple of years.

We started by training one deep autoencoder using public genuine face databases that models the typical spatial correlations between the pixels of real faces and simultaneously removes the high frequency components that correspond to the "fingerprints" of the models used to generate synthetic images. In test time, the autoencoder was fed only with synthetic face images to produce manipulated versions, whose properties were deliberately changed for spoofing fake detection systems.

In the empirical validation of our approach, we used various well known face datasets, coming out with three major conclusions about the performance of the state-of-the-art fake detection methods: *i)* the existing fake systems attain almost perfect performance when the evaluation data is derived from the same source used in the training phase, which suggests that these systems have actually learned the GAN "fingerprints" from the training fakes generated with GANs; *ii)* the observed fake detection performance decreases substantially (over one order of magnitude) when the fake detection is exposed to data from unseen databases, and over seven times in case of substantially reduced image resolution; and *iii)* the accuracy of the existing fake detection methods also drops significantly when analysing synthetic data manipulated by GANprintR.

In summary, our experiments suggest that the existing facial fake detection methods still have a poor generalisation capability and are highly susceptible to - even simple - image transformation manipulations, such as downsizing, image compression or others similar to the one proposed in this work. While loss of resolution may not be particularly concerning in terms of the potential misuse of the data, it is important to note that our approach is capable of confounding detection methods, while maintaining a high visual similarity with the original image.

Having shown some of the limitations of the state of the art in face manipulation detection, in future work we can harden such face manipulation detectors by exploiting our improved fakes. Additionally, further works may study: *i)* how improved fakes obtained in similar ways as GANprintR can jeopardize other kinds of sensitive data (e.g., other popular biometrics like fingerprint [50], iris [51], or behavioral traits [52], [53]), *ii)* how to improve the security of systems dealing with other kinds of sensitive data, and finally *iii)* best ways to combine multiple manipulation detectors [54] in a proper way to deal with the growing sophistication of fakes.

Acknowledgment

We gratefully acknowledge the donation of the NVIDIA Titan X GPU used for this research made by NVIDIA Corporation. Ruben Tolosana is supported by Consejería de Educación, Juventud y Deporte de la Comunidad de Madrid y Fondo Social Europeo.

References

[1] R. Cellan-Jones, “Deepfake videos double in nine months,” 2019. [Online]. Available: https://www.bbc.com/news/technology-49961089

[2] A. Hadid, N. Evans, S. Marcel, and J. Fierrez, “Biometrics systems under spoofing attack: An evaluation methodology and lessons learned,” *IEEE Signal Process. Mag.*, vol. 32, no. 5, pp. 20–30, Sep. 2015.

[3] J. Hernandez-Ortega, J. Fierrez, A. Morales, and J. Galbally, “Introduction to face presentation attack detection,” in *Handbook of Biometric Anti-Spoofing: Presentation Attack Detection*, S. Marcel, M. S. Nixon, J. Fierrez, and N. Evans, Eds., Berlin, Germany: Springer, 2019, pp. 187–206.

[4] J. Galbally, S. Marcel, and J. Fierrez, “Biometric anti-spoofing methods: A survey in face recognition,” *IEEE Access*, vol. 2, pp. 1530–1552, 2014.

[5] “ZAO,” 2019. [Online]. Available: https://apps.apple.com/cn/app/id1465199127

[6] “FaceApp,” 2017. [Online]. Available: https://apps.apple.com/us/app/faceapp-ai-face-editor/id1180884341

[7] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, “DeepFakes and beyond: A survey of face manipulation and fake detection,” *Inf. Fusion*, 2020.

[8] L. Verdoliva, “Media forensics and DeepFakes: An overview,” *IEEE J. Sel. Topics Signal Proc.*, 2020, *arXiv:2001.06564*.

[9] H. Dang, F. Liu, J. Stehouwer, X. Liu, and A. Jain, “On the detection of digital face manipulation,” in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit.*, 2020, pp. 5781–5790.

[10] S. McCloskey and M. Albright, “Detecting GAN-generated imagery using color cues,” 2018, *arXiv:1812.08247*.

[11] N. Yu, L. Davis, and M. Fritz, “Attributing fake images to GANs: Analyzing fingerprints in generated images,” in *Proc. IEEE/CVF Int. Conf. Comput. Vision*, 2019, pp. 7556–7566.

[12] F. Marra, D. Gragnaniello, D. Cozzolino, and L. Verdoliva, “Detection of GAN-generated fake images over social networks,” in *Proc. IEEE Conf. Multimedia Inf. Process. Retrieval*, 2018, pp. 384–389.

[13] R. Wang, L. Ma, F. Juefei-Xu, X. Xie, J. Wang, and Y. Liu, “FakeSpotter: A simple baseline for spotting AI-synthesized fake faces,” 2019, *arXiv:1909.06122*.

[14] X. Yang, Y. Li, and S. Lyu, “Exposing deep fakes using inconsistent head poses,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 8261–8265.

[15] F. Matern, C. Riess, and M. Stamminger, “Exploiting visual artifacts to expose deepfakes and face manipulations,” in *Proc. IEEE Winter Appl. Comput. Vision Workshops*, 2019, pp. 83–92.

[16] P. He, H. Li, and H. Wang, “Detection of fake images via the ensemble of deep representations from multi color spaces,” in *Proc. IEEE Int. Conf. Image Process.*, 2019, pp. 2299–2303.

[17] S. Y. Wang, O. Wang, A. Owens, R. Zhang, and A. A. Efros, “Detecting photoshopped faces by scripting photoshop,” in *Proc. IEEE/CVF Int. Conf. Comput. Vision*, 2019, pp. 10071–10080.

[18] I. Goodfellow *et al.*, “Generative Adversarial Nets,” in *Proc. Advances Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[19] National Institute of Standards and Technology. “Media forensics challenge,” 2018. [Online]. Available: https://www.nist.gov/itl/iad/mig/media-forensics-challenge-2018

[20] O. Parkhi, A. Vedaldi, and A. Zisserman, “Deep face recognition,” in *Proc. Brit. Mach. Vision Conf.*, 2015, pp. 41.1–41.12.

[21] B. Amos, B. Ludwiczuk, and M. Satyanarayanan, “OpenFace: A general-purpose face recognition library with mobile applications,” *CMU School Comput. Sci.*, 2016, pp. 1–18.

[22] F. Schroff, D. Kalenichenko, and J. Philbin, “FaceNet: A unified embedding for face recognition and clustering,” in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 815–823.

[23] “CelebA-HQ,” 2018. [Online]. Available: https://drive.google.com/drive/folders/0B4qLcYyJmiz0TXY1NG02bzZVRGs

[24] T. Karras, S. Laine, and T. Aila, “A style-based generator architecture for generative adversarial networks,” in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 4396–4405.

[25] Y. Shen, J. Gu, X. Tang, and B. Zhou, “Interpreting the latent space of GANs for semantic face editing,” in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit.*, 2020, pp. 9243–9252.

[26] Z. Liu, P. Luo, X. Wang, and X. Tang, “Deep learning face attributes in the wild,” in *Proc. IEEE/CVF Int. Conf. Comput. Vision*, 2015, pp. 3730–3738.

[27] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, “FaceForensics++: Learning to detect manipulated facial images,” in *Proc. IEEE/CVF Int. Conf. Comput. Vision*, 2019, pp. 1–11.

[28] T. Karras, T. Aila, S. Laine, and J. Lehtinen, “Progressive growing of GANs for improved quality, stability, and variation,” in *Proc. Int. Conf. Learn. Representations*, 2018, pp. 1–26.

[29] Adjust and exaggerate facial features, “Adobe Photoshop,” 2016. [Online]. Available: https://helpx.adobe.com/photoshop/how-to/face-aware-liquify.html

[30] M. Albright and S. McCloskey, “Source generator attribution via inversion?” in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit. Workshops*, 2019, pp. 1–8.

[31] F. Marra, D. Gragnaniello, L. Verdoliva, and G. Poggi, “Do GANs leave artificial fingerprints?” in *Proc. IEEE Conf. Multimedia Inf. Process. Retrieval*, 2019, pp. 506–511.

[32] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, “Spectral normalization for generative adversarial networks,” in *Proc. Int. Conf. Learn. Representations*, 2018, pp. 1–26.

[33] M. Bellemare *et al.*, “The cramer distance as a solution to biased wasserstein gradients,” 2017, *arXiv:1705.10743*.

[34] M. Binkowski, D. Sutherland, M. Arbel, and A. Gretton, “Demystifying MMD GANs,” in *Proc. Int. Conf. Learn. Representations*, 2018, pp. 1–36.

[35] X. Zhang, S. Karaman, and S. Chang, “Detecting and simulating artifacts in GAN fake images,” in *Proc. IEEE Int. Workshop Inf. Forensics Secur.*, 2019, pp. 1–6.

[36] M. Huh, A. Liu, A. Owens, and A. A. Efros, “Fighting fake news: Image splice detection via learned self-consistency,” in *Proc. Eur. Conf. Comput. Vision*, 2018.

[37] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, “Learning rich features for image manipulation detection,” in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 1053–1061.

[38] L. Nataraj *et al.*, “Detecting GAN generated fake images using co-occurrence matrices,” *Electron. Imag.*, vol. 5, no. 5, pp. 1–7, 2019.

[39] D. Yi, Z. Lei, S. Liao, and S. Li, “Learning face representation from scratch,” 2014, *arXiv:1411.7923*.

[40] Q. Cao, L. Shen, W. Xie, O. Parkhi, and A. Zisserman, “VGGFace2: A dataset for recognising faces across pose and age,” in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2018, pp. 67–74.

[41] “Flickr-Faces-HQ Dataset (FFHQ),” 2019. [Online]. Available: https://github.com/NVlabs/ffhq-dataset

[42] “100,000 Faces Generated by AI,” 2018. [Online]. Available: https://generated.photos/

[43] V. Kazemi and J. Sullivan, “One millisecond face alignment with an ensemble of regression trees,” in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit.*, 2014, pp. 1867–1874.

[44] F. Chollet, “Xception: Deep learning with depthwise separable convolutions,” in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 1800–1807.

[45] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. C. Ferrer, “The deepfake detection challenge (DFDC) preview dataset,” 2019, *arXiv:1910.08854*.

[46] 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/CVF Conf. Comput. Vision Pattern Recognit.*, 2009, pp. 248–255.

[47] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, “MS-Celeb-1M: A dataset and benchmark for large scale face recognition,” in *Proc. Eur. Conf. Comput. Vision*, 2016, pp. 87–102.

[48] P. Korshunov and S. Marcel, “DeepFakes: A new threat to face recognition? Assessment and detection,” 2018, *arXiv:1812.08685*.

[49] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in *Proc. IEEE/CVF Int. Conf. Comput. Vision*, 2017, pp. 5967–5976.

[50] R. Tolosana, M. Gomez-Barrero, C. Busch, and J. Ortega-Garcia, “Biometric presentation attack detection: Beyond the visible spectrum,” *IEEE Trans. Inf. Forensics Secur.*, vol. 15, pp. 1261–1275, 2020.

[51] H. Proenca and J. C. Neves, “Segmentation-less and non-holistic deep-learning frameworks for iris recognition,” in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit. Workshops*, 2019, pp. 2296–2305.

[52] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, and J. Ortega-Garcia, “BioTouchPass2: Touchscreen password biometrics using time-aligned recurrent neural networks,” *IEEE Trans. Inf. Forensics Secur.*, vol. 15, pp. 2616–2628, 2020.

[53] A. Morales *et al.*, “Keystroke biometrics in response to fake news propagation in a global pandemic,” in *Proc. IEEE Comput. Softw. Appl. Conf. Workshops*, 2020.

[54] J. Fierrez, A. Morales, R. Vera-Rodriguez, and D. Camacho, “Multiple classifiers in biometrics. Part 2: Trends and challenges,” *Inf. Fusion*, vol. 44, pp. 103–112, 2018.

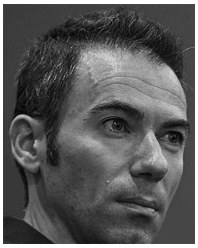
João C. Neves received the B.Sc., M.Sc., and Ph.D. degrees in computer science from the University of Beira Interior, Covilhã, Portugal, in 2011, 2013, and 2018, respectively. He is currently an Assistant Professor with the University of Beira Interior. His current research interests broadly include computer vision and pattern recognition, with a particular focus on biometrics and surveillance. He is author of several publications and also collaborates as a reviewer in many different high-impact conferences (e.g., WAVC, IJCB, ACM MM, etc.) and journals (e.g., IEEE TMI, TIFS, TCYB, TCSVT, etc.).

Ruben Tolosana received the M.Sc. degree in telecommunication engineering, and the Ph.D. degree in computer and telecommunication engineering from Universidad Autonoma de Madrid, Madrid, Spain, in 2014 and 2019, respectively. In April 2014, he joined the Biometrics and Data Pattern Analytics - BiDA Lab, Universidad Autonoma de Madrid, where he is currently collaborating as a Postdoctoral Researcher. Since then, he has been granted with several awards, such as the FPU research fellowship from Spanish MECD (2015), and the European Biometrics Industry Award (2018). His research interests are mainly focused on signal and image processing, pattern recognition, deep learning, and biometrics, particularly in the areas of handwriting and handwritten signature. He is author of several publications and also collaborates as a reviewer in many different high-impact conferences (e.g., ICDAR, ICB, BTAS, EUSIPCO, etc.) and journals (e.g., IEEE TPAMI, TIFS, TCYB, TIP, ACM Computing Surveys, etc.). Finally, he has participated in several National and European projects focused on the deployment of biometric security through the world.

Ruben Vera-Rodriguez received the M.Sc. degree in telecommunications engineering from Universidad de Sevilla, Seville, Spain, in 2006, and the Ph.D. degree in electrical and electronic engineering from Swansea University, Swansea, U.K., in 2010. Since 2010, he has been affiliated with the Biometric Recognition Group, Universidad Autonoma de Madrid, Madrid, Spain, where he is currently an Associate Professor since 2018. His current research interests include signal and image processing, pattern recognition, and biometrics, with emphasis on signature, face, gait verification and forensic applications of biometrics. He is actively involved in several National and European projects focused on biometrics. Dr. Vera-Rodriguez has been Program Chair for the IEEE 51st International Carnahan Conference on Security and Technology (ICCST) in 2017; and the 23rd Iberoamerican Congress on Pattern Recognition (CIARP 2018) in 2018.

Vasco Lopes received the B.Sc. and M.Sc. degrees in computer science and engineering in 2017 and 2019, respectively, from the University of Beira Interior, Covilhã, Portugal, where he is currently working toward the a Ph.D. degree in the field of Artificial Intelligence, with focus on computer vision. His current research interests broadly include computer vision, robotics, and artificial intelligence. Mr. Lopes was the recipient of the APRP Best Dissertation in Pattern Recognition 2019 Award.

Hugo Proença received the B.Sc., M.Sc. and Ph.D. degrees in 2001, 2004, and 2007, respectively. He is currently an Associate Professor with the Department of Computer Science, University of Beira Interior, Covilhã, Portugal and has been researching mainly about biometrics and visual-surveillance. He is the Coordinating Editor for the IEEE Biometrics Council Newsletter and the Area Editor (ocular biometrics) for the IEEE Biometrics Compendium Journal. He is a Member of the Editorial Boards for the *Image and Vision Computing* and *International Journal of Biometrics* and served as a Guest Editor of special issues of the *Pattern Recognition Letters*, *Image and Vision Computing* and *Signal, Image and Video Processing* journals.



Julian Fierrez received the M.Sc. and Ph.D. degrees in telecommunications engineering from the Universidad Politecnica de Madrid, Madrid, Spain, in 2001 and 2006, respectively. Since 2004, he has been with Universidad Autonoma de Madrid, Madrid, Spain, where he is currently an Associate Professor. From 2007 to 2009, he was a Visiting Researcher with Michigan State University, USA, under a Marie Curie postdoc. His research interests include signal and image processing, HCI, responsible AI, and biometrics for security and human behavior analysis. He is actively involved in large EU projects in these topics (e.g., TABULA RASA and BEAT in the past, now IDEA-FAST and TRESPASS-ETN), and has attracted notable impact for his research. He was the recipient of a number of distinctions, including the EAB Industry Award 2006, the EURASIP Best Ph.D. Award 2012, and the 2017 IAPR Young Biometrics Investigator Award. He has received Best Paper Awards at ICB and ICPR. He is an Associate Editor for the IEEE Transactions on Information Forensics and Security and the IEEE Transactions on Image Processing. He is Member of the ELLIS Society.