Chapter 8
GAN Fingerprints in Face Image Synthesis

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

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. This chapter is focused on the analysis of GAN fingerprints in face image synthesis. In particular, it covers an in-depth literature analysis of state-of-the-art detection approaches for the entire face synthesis manipulation. It also describes a recent approach to spoof fake detectors based on a GAN-fingerprint Removal autoencoder (GANprintR). A thorough experimental framework is included in the chapter, highlighting *(i)* the potential of GANprintR to spoof fake detectors, and *(ii)* the poor generalisation capability of current fake detectors.

8.1 Introduction

Images[1] and videos containing fake facial information obtained by digital manipulation have recently become a great public concern (Cellan-Jones 2019). Up until the advent of DeepFakes a few years ago, the number and realism of digitally manip- AQ1

[1]The present chapter is an adaptation from the following article: Neves et al. (2020). DOI: http://dx.doi.org/10.1109/JSTSP.2020.3007250.

J. C. Neves · R. Tolosana · R. Vera-Rodriguez · V. Lopes · H. Proença · J. Fierrez (✉)
NOVA LINCS, Universidade da Beira Interior, Covilha, Portugal
e-mail: julian.fierrez@uam.es

BiDA-Lab, Universidad Autonoma de Madrid, Madrid, Spain

H. T. Sencar et al. (eds.), *Multimedia Forensics*, Advances in Computer Vision and Pattern Recognition, https://doi.org/10.1007/978-981-16-7621-5_8

ulated fake facial contents were very 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. The scientific communities of biometrics and security in the past decade paid some attention in understanding and protecting against those limited threats around face biometrics (Hadid et al. 2015), with special attention to 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.) (Hernandez-Ortega et al. 2019; Galbally et al. 2014).

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 (ZAO 2019; FaceApp 2017).

The current methods to generate digital fake face content can be categorised into four different groups, regarding the level of manipulation (Tolosana et al. 2020c; Verdoliva 2020): *(i)* entire face synthesis, *(ii)* face identity swap, *(iii)* facial attribute manipulation and *(iv)* facial expression manipulation.

In this chapter, we focus on the entire face synthesis manipulation, where a machine learning model, typically based on Generative Adversarial Networks (GANs) (Goodfellow et al. 2014), 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.

This chapter covers the following aspects in the topic of GAN Fingerprints:

- 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 databases used for developing and evaluating these systems, and the main results achieved by them.
- An approach to spoof state-of-the-art facial manipulation detection systems, while keeping the visual quality of the resulting images. Figure 8.1 graphically summarises the approach presented in Neves et al. (2020) based on a GAN-fingerprint Removal autoencoder (GANprintR).
- 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.

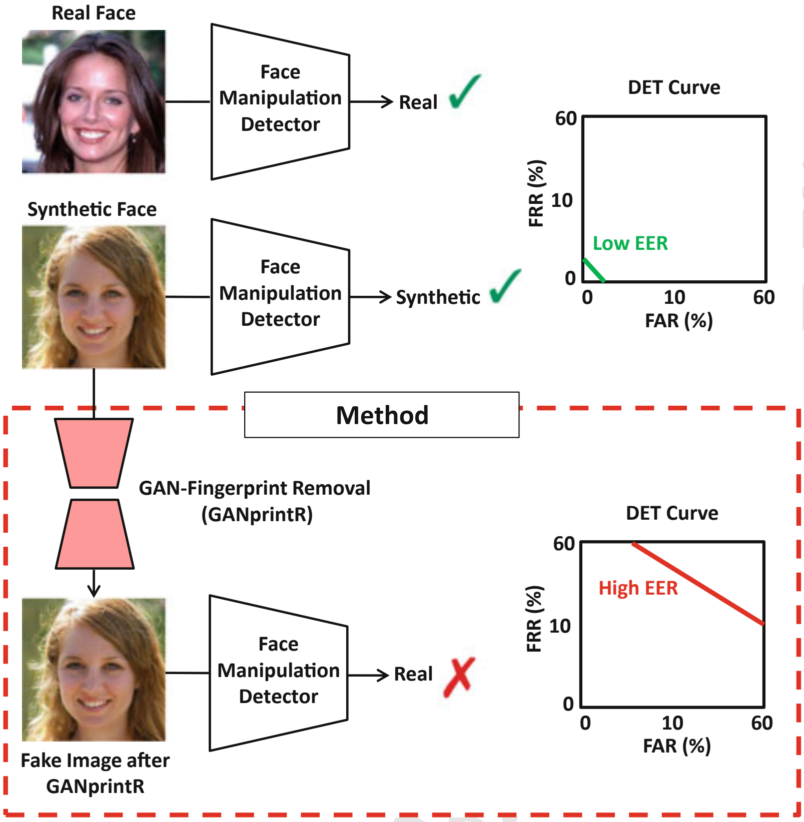


Fig. 8.1 Architecture of the GAN-fingerprint removal approach. In general, the state-of-the-art face manipulation detectors can easily distinguish between real and synthetic fake images. This usually happens due to the existence and exploitation by those detectors of GAN "fingerprints" produced during the generation of synthetic images. The GANprintR approach proposed in Neves et al. (2020) aims 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

- A recent database named iFakeFaceDB,[2] resulting from the application of the GANprintR approach to already very realistic synthetic images.

The remainder of the chapter is organised as follows. Section 8.2 summarises the state of the art on the exploitation of GAN fingerprints for the detection of entire face synthesis manipulation. Section 8.3 explains the GAN-fingerprint removal approach (GANprintR) presented in Neves et al. (2020). Section 8.4 summarises the key features of the real and fake databases considered in the experimental assessment of this type of facial manipulation. Sections 8.5 and 8.6 describe the experimental setup and results achieved, respectively. Finally, Sect. 8.7 draws the final conclusions and points out some lines for future work.

[2] https://github.com/socialabubi/iFakeFaceDB.

8.2 Related Work

Contrary to popular belief, image manipulation dates back to the dawn of photography. Nevertheless, image manipulation only became particularly important after the rise of digital photography, due to the use of image processing techniques or low-cost image editing software. As a consequence, in the last decades the research community devised several strategies for assuring authenticity of digital data. In addition, digital image tampering still required some level of expertise to deceive the humans' eye, and both factors helped reducing significantly the use of manipulated content for malicious purposes. However, after the proposal of Generative Adversarial Networks (Goodfellow et al. 2014), the possibility of synthesising realistic digital content became possible. Among the four possible levels of face manipulation, this chapter focuses on the entire face synthesis manipulation, particularly on the problem of distinguishing between real and fake facial images.

Typically, synthetic face detection methods rely on the "fingerprints" caused by the generation process. According to the type of fingerprints used, each approach can be broadly divided into three categories: *(i)* methods based on visual artifacts; *(ii)* methods based on frequency analysis; and *(iii)* learning-based approaches for automatic fingerprint estimation. Table 8.1 provides a comparison of the state-of-the-art synthetic face detection methods.

The following sections describe the state-of-the-art techniques for synthetic data generation and review the state-of-the-art methods capable of detecting synthetic face imagery according to the taxonomy described above.

8.2.1 Generative Adversarial Networks

Proposed by Goodfellow et al. (2014), GANs are a novel generative concept, composed of two neural networks contesting each other in the form of a competition. A generator learns to generate instances that resemble the training data, while a discriminator learns to distinguish between the real and the generated images, while serving the goal of penalising the generator. The goal is to have a generator that can learn how to generate plausible images that can fool the discriminator. While at the beginning, GANs were only capable of producing low-resolution images of faces with some notorious visual artifacts, in the last years several techniques have emerged for synthesising highly realistic content (including BigGAN Brock et al. 2019, CycleGAN Zhu et al. 2017, GauGAN Park et al. 2019, ProGAN Karras et al. 2018, StarGAN Choi et al. 2018, StyleGAN Karras et al. 2019, and StyleGAN2 Karras et al. 2020) that even humans cannot distinguish from the real ones. Next, we review the state-of-the-art approaches specifically devised for detecting a entire face synthesis manipulation.

Table 8.1 Comparison of the state-of-the-art synthetic face detection methods

Study	Features	Classifiers	Best performance	Databases
Visual artifacts				
McCloskey and Albright (2018)	Colour Histogram	SVM	AUC = 70%	NIST MFC2018
Matern et al. (2019)	Eye Colour	K-NN	AUC = 85.2%	Real: CelebA Fake: Own Database (PGGAN)
Yang et al. (2019)	Head Pose	SVM	AUC = 89%	Real: UADFV/DARPA MediFor Fake: UADFV/DARPA MediFor
He et al. (2019)	Colour-related	Random Forest	Acc. = 99%	Real: CelebA Fake: Own Database (PGGAN)
Li et al. (2020)	Correlation Between Adjacent Pixels in Multiple Colour Channels	–	Acc. = 91.87	Real: FFHQ/LFW/LSUN/FFHQ Fake: Own Database (ProGAN, StyGAN, BigGAN, CocoGAN, DCGAN and WGAN-GP)
Hu et al. (2020)	Difference Between the Two Corneal Specular Highlights	Rule-based	AUC = 94%	Real: FFHQ Fake: Own Database (StyleGan2)
High-frequency information				
Yu et al. (2018)	GAN Fingerprint	Rule-based	Acc. = 99.50%	Real: CelebA Fake: Own Database
Wang et al. (2020b)	CNN Neuron Behaviour	SVM	Acc. = 84.78%	Real: CelebA-HQ/FFHQ Fake: Own Database
Stehouwer et al. (2020)	Image-related	CNN + Attention	EER = 0.05%	Real: CelebA/FFHQ/FaceForensics++ Fake: Own Database
Marra et al. (2019a)	GAN Fingerprint	Rule-based	AUC = 99.9%	Real: RAISE Fake: Own Database (Cycle-GAN, ProGAN, and Star-GAN)
Albright et al. (2019)	GAN Fingerprint	Rule-based	Acc. = 98.33%	Real: MNIST/CelebA Fake: Own Database (ProGAN, SAGAN, SNGAN)
Guarnera et al. (2020)	Local Pixel Correlations	K-NN	Acc. = 99.81%	Real: CelebA Fake: Own Database (StarGAN, StyleGAN, StyleGAN2, GDWCT, AttGAN)

513475_1_En_8_Chapter ✓ TYPESET ☐ DISK ☐ LE ✓ CP Disp.:13/11/2021 Pages: xxx Layout: T1-Standard

Table 8.1 (continued)

Study	Features	Classifiers	Best performance	Databases
Visual artifacts				
Zhang et al. (2019)	Image Spectrum	CNN	Acc. = 97.2%	CycleGAN/AutoGAN
Durall et al. (2020)	Frequency features extracted from DFT	SVM	Acc. = 90%	Real: Own Database (CelebA, FFHQ) Fake: Own Database (100 K, StyleGan)
Frank et al. (2020)	Frequency features extracted from DCT	Ridge-regression	Acc. = 100%	Real: FFHQ Fake: Own Database (StyleGAN)
Bonettini et al. (2020)	Distribution of the quantized coefficients of the DCT	Random Forest	Acc. = 99.83%	GAN-generated (Marra et al. 2019b) (CycleGAN, ProGAN)
Learning-based				
Marra et al. (2018)	Image-related	CNN	Acc. = 95.07%	Real: Own Database(CycleGAN) Fake: Own Database(CycleGAN)
Hsu et al. (2020)	Raw Image	CNN	Precision = 88 Recall = 87.32	Real: CelebA Fake: Own Database (DCGAN, WGAP, WGAN-GP, LSGAN, PGGAN)
Marra et al. (2019c)	Raw Image Using Incremental Learning Strategy	CNN	Acc. = 99.37%	Real: CelebA-HQ Fake: DoGANS (CycleGAN, ProGAN, Glow, StarGAN)
Xuan et al. (2019)	Pre-processed Image Using Blur or Noise in Training	CNN	Acc. = 95.45%	Real: CelebA-HQ Fake: Own Database (DC-GAN, WGAN-GP, PGGAN)
Wang et al. (2020a)	Raw Image	CNN	mAP = 93	Own Database (using 11 synthesis models)
Hsu et al. (2020)	Raw Image	CNN	Precision = 96.76 Recall = 90.56	Real: CelebA Fake: Own Database (DCGAN, WGAP, WGAN-GP, LSGAN, PGGAN)
Nataraj et al. (2020)	Co-occurrence matrix of each colour channel (RGB)	CNN	Acc. = 87.96%	Own Database (ProGAN, StarGAN, GlowGAN, StyleGAN2)

Table 8.1 (continued)

Study	Features	Classifiers	Best performance	Databases
isual artifacts				
Goebel et al. (2020)	Co-occurrence matrix of each colour channel (RGB)	CNN	Acc. = 98.17%	Own Database (StarGAN, CycleGAN, ProGAN, Spade, StyleGAN)
Bani et al. (2020)	Co-occurrence matrix of each colour channel (RGB) and for each colour channels pairs	CNN	Acc. = 99.70%	Real: FFHQ Fake: Own Database (StyleGAN2)
Hulzebosch et al. (2020)	Pre-processed Image Using Colour Transformations, Co-occurrence Matrices or High-pass Filters	CNN	Acc. = 99.9%	Real: CelebA-HQ/FFHQ Fake: Own Database (StarGAN, GLOW, ProGAN, StyleGAN)
Liu et al. (2020)	Global Texture Features captured by "Gram-Block" (extra layer)	CNN	Acc. = 95.51%	Real: CelebA-HQ/FFHQ Fake: Own Database (StyleGAN, PGGAN, DCGAN, DRAGAN, StarGAN)
Yu et al. (2020a)	Channel Differences, Image Spectrum	CNN	Acc. = 99.41%	Real: FFHQ Fake: Own Database (StyleGAN, StyleGAN2)

513475_1_En_8_Chapter ✓ TYPESET DISK LE ✓ CP Disp.:13/11/2021 Pages: xxx Layout: T1-Standard

8.2.2 GAN Detection Techniques

As denoted before, the images generated by the initial versions of GANs exhibited several visual artifacts, including distinct eye colour, holes in the face, deformed teeth, among others. For this reason, several approaches attempted to leverage these traits for detecting face manipulations (Matern et al. 2019; Yang et al. 2019; Hu et al. 2020). Matern et al. (2019) extracted several geometric facial features which were then fed to a Support Vector Machine (SVM) classifier to distinguish between real and synthetic face images. Yang et al. (2019) 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. As the remaining artifacts became less noticeable, researchers focused on more subtle features of the face, as in Hu et al. (2020), where synthetic face detection was performed by analysing the difference between the two corneal specular highlights. Other visual artifact typically exploited is the probability distribution of colour channels. McCloskey (McCloskey and Albright 2018) hypothesised that the colour is markedly different between real camera images and fake synthesis images, and proposed a detection system based on the colour histogram and a linear SVM. In He et al. (2019), the authors exploited different colour channels (YCbCr, HSV and Lab) to extract from a CNN different deep representations, which were subsequently fed to a Random Forest classifier for distinguishing between real and synthetic data. Li et al. (2020) observed that it is easier to spot the differences between real and GAN-generated data in non-RGB colour spaces, since GANs are trained for producing content in RGB channels.

As the quality and realism of synthetic data improved, visual artifacts started to become ineffectual, which in turn fostered researchers to explore digital forensic techniques for the problem of synthetic data detection. Each camera sensor leaves a unique and stable mark on each acquired photo, denoted as the photo-response non-uniformity (PRNU) pattern (Lukás et al. 2006). This mark is usually denoted as the camera fingerprint, which inspired researchers to detect the presence of similar patterns in images synthesised by GANs. These approaches usually define the GAN fingerprint as a high-frequency signal available in the image. Marra et al. (2019a) defined GAN fingerprint as the high-level image information obtained by subtracting the image from its corresponding denoised version. Yu et al. (2018) improved (Marra et al. 2019a) by subtracting from the original image the corresponding reconstructed version obtained from an autoencoder, which was tuned based on the discriminability of the fingerprints inferred by this process. 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 gives the probability of the image being produced by a specific model. Their proposed approach was tested using real faces from CelebA database (Liu et al. 2015) and synthetic faces created through different GAN approaches (PGGAN Karras et al. 2018, SNGAN Miyato et al. 2018, CramerGAN Bellemare et al. 2017, and MMDGAN Binkowski et al. 2018), achieving a final accuracy of 99.50% for the best performance. Later, they extended their approach Yu et al. (2020b) by proposing a novel strategy for the training of the generative model such that the fingerprints can be controlled by the user, and easily decoded from a synthetic image, allowing to solve the problem of source attribution, i.e. identifying the model that generated the image. In Albright and McCloskey (2019), the authors proposed an alternative to Yu et al. (2018) by replacing the autoencoder by an inverted GAN capable of reconstructing an image based on the attributes inferred from the original image. Zhang et al. (2019) proposed the use of the up-sampling artifact in the frequency domain as a discriminative feature for distinguishing veridical and synthetic data. Frank et al. (2020) reported similar conclusions regarding the discriminability of the frequency space of GAN-generated images. They relied on the Discrete Cosine Transform (DCT) for extracting features from either real and fake images, in order to train a linear classifier. Durall et al. (2020) found out that upconvolution or transposed convolution layers of GAN architectures are not capable of reproducing the spectral distribution of natural images. Based on this finding, they showed that generated face images can be easily identified by training a SVM with the features extracted with the Discrete Fourier Transform (DFT). Guarnera et al. (2020) used pixel correlation as a GAN fingerprint, since they noticed that the correlation of pixels in synthetic images are exclusively dependent on the operations performed by all the layers present in the GAN which generate it. Their proposed approach was tested using fake images generated by several GAN architectures (AttGAN, GDWCT, StarGAN, StyleGAN and StyleGAN2).

A distinct family of methods adopts a data-driven strategy for the problem of detecting GAN-generated imagery. In this strategy, a standard image classifier, typically a Convolutional Neural Network (CNN), is trained directly with raw images or through a modified version of them (Barni et al. 2020; Hsu et al. 2020). Marra et al. (2018) carried out a study about the classification accuracy of different CNN architectures when fed with raw images. It was observed that, in spite almost ideal performance was obtained, the performance decreased significantly when compressed images were used in the test set. Later, the authors proposed a strategy based on incremental learning for addressing this problem and the generalisation to unseen datasets (Marra et al. 2019c). Inspired by the forensic analysis of image manipulation (Cozzolino et al. 2014), Nataraj et al. (2019a) proposed a detection system based on a combination of pixel co-occurrence matrices and CNNs. Their proposed approach was initially tested in a database of various objects and scenes created through CycleGAN (Zhu et al. 2017). Besides, the authors performed an interesting analysis to see the robustness of the proposed approach against fake images created through different GAN architectures (CycleGAN vs. StarGAN), with good generalisation results. This idea was later improved in Goebel et al. (2020) and Barni et al. (2020).

The above studies show that a simple CNN is able to easily distinguish between real and synthetic data generated from specific GAN architectures, but is not capable of maintaining the same performance in data originated from GAN architectures not seen during training or even in data altered by image filtering operations. For this reason, Xuan et al. (2019) used an image pre-processing step in the training stage to remove artifacts of a specific GAN architecture. The same idea was exploited in Hulzebosch et al. (2020) to improve the accuracy in real-world scenarios, where the particularities of the data (e.g. image compression) and the generator architecture

513475_1_En_8_Chapter ☑ TYPESET ☐ DISK ☐ LE ☑ CP Disp.:13/11/2021 Pages: xxx Layout: T1-Standard

are not known. Liu et al. (2020) observed that the texture of fake faces is substantially different from the real ones. Based on this observation, the authors devised a novel block to be added to the backbone of a CNN, the Gram-Block, which is capable of extracting global image texture features and improve the generalisation of the model against data generated by GAN architectures not used during training. Similarly, Yu et al. (2020a) introduced a novel convolution operator intended for separately processing the low- and high-frequency information of the image, improving the capability to detect the patterns of synthetic data available in the high-frequency band of the images. Finally, Wang et al. (2020a) studied the topic of generalisation to unseen datasets. For this, they collected a dataset consisting of fake images generated by 11 different CNN-based image generator models and concluded that the correct combination of pre-processing and data augmentation techniques allows a standard image classifier to generalise to unseen dataset even when trained with data obtained from a single GAN architecture.

To summarise this section, we conclude that state-of-the-art automatic detection systems against face synthesis manipulation have excellent performance, mostly because they are able to learn the GAN fingerprints present in the images. However, it is also clear that the dependence on the model fingerprint affects the generability and the reliability of the model, e.g. when presented with adversarial attacks (Gandhi and Jain 2020).

8.3 GAN Fingerprint Removal: GANprintR

GANprintR was originally presented in Neves et al. (2020) and 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 (Marra et al. 2019a), we hypothesised that their removal could be performed by an autoencoder, which acts as a non-linear low-pass filter. We claimed 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 autoencoder comprises two distinct networks, encoder ψ and decoder γ:

$$\begin{aligned} \psi &: X \mapsto l \\ \gamma &: l \mapsto X', \end{aligned} \tag{8.1}$$

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 $\mathcal{L}_{\psi,\gamma}(X, X') = ||X - X'||^2$ over a development dataset following an iterative learning strategy.

513475_1_En_8_Chapter ☑ TYPESET ☐ DISK ☐ LE ☑ CP Disp.:13/11/2021 Pages: xxx Layout: T1-Standard

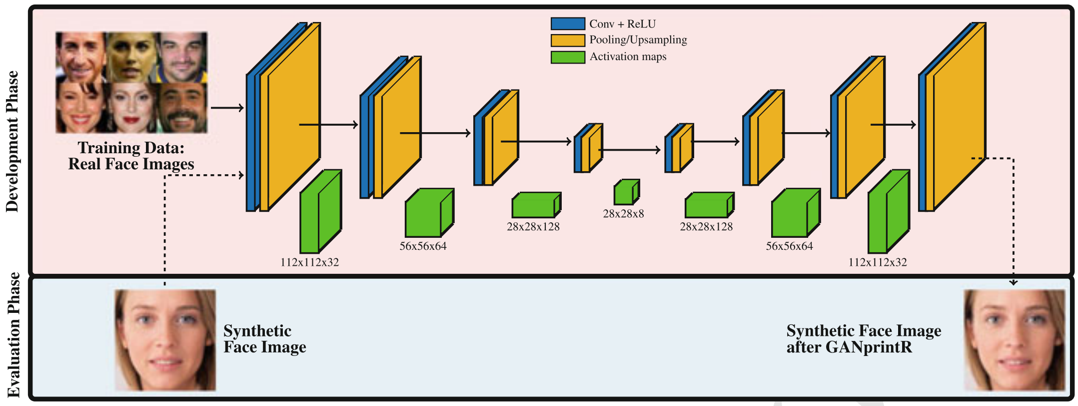


Fig. 8.2 GAN-fingerprint Removal module (GANprintR) based on a convolutional AutoEncoder (AE). The AE is trained using only real face images from the development dataset. In the evaluation stage, once the autoencoder is trained, we can pass synthetic face images through it to provide them with additional naturalness, in this way removing the GAN-fingerprint information that may be present in the initial fakes

As result, when $\mathcal{L}$ 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, $\mathcal{L}$ will increase and ψ will be forced to encode in l only the most representative information of X. We claimed that this kind of autoencoder acts as a GAN-fingerprint removal system.

Figure 8.2 describes the GANprintR architecture 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 $28 \times 28 \times 8$, 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.

8.4 Databases

Four different public databases and one generated are considered in the experimental framework of this chapter. Figure 8.3 shows some examples of each database. We now summarise the most important features.

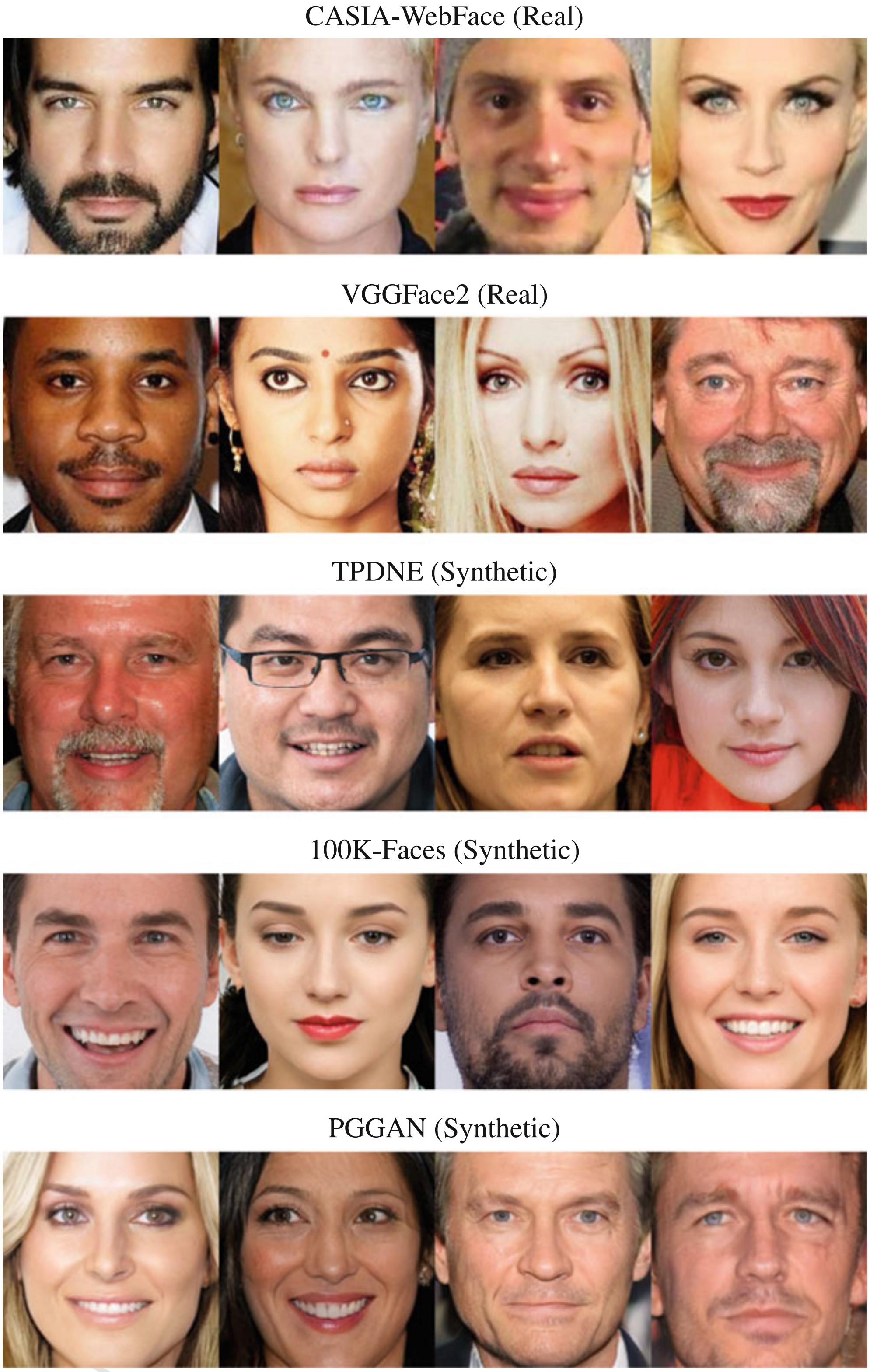


Fig. 8.3 Examples of the databases considered in the experiments of this chapter after applying the pre-processing stage described in Sect. 8.5.1

8.4.1 Real Face Images

- *CASIA-WebFace:* 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.
- *VGGFace2:* 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).

8.4.2 Synthetic Face Images

- *TPDNE:* this database comprises 150,000 unique faces, collected from the website.[3] Synthetic images are based on the recent StyleGAN approach (Karras et al. 2019) trained with FFHQ database (Flickr-Faces-HQ 2019).
- *100K-Faces:* this database contains 100,000 synthetic images generated using StyleGAN (Karras et al. 2019). In this database the StyleGAN network was trained using around 29,000 photos of 69 different models, producing face images with a flat background.
- *PGGAN:* 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.

8.5 Experimental Setup

This section describes the details of the experimental setup followed in the experimental framework of this chapter.

8.5.1 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 chapter:

- *Background*: this is a clearly distinctive aspect among real and synthetic face images as different acquisition conditions are considered in each database.

[3] https://thispersondoesnotexist.com.

- *Head pose*: images generated by GANs hardly ever produce high variation from the frontal pose (Dang et al. 2020), 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 Kazemi and Sullivan (2014). 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 the experimental framework of this chapter, 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. Figure 8.3 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 the GANprintR approach.

8.5.2 *Facial Manipulation Detection Systems*

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

(1) *XceptionNet* (Chollet 2017): this network was selected, essentially because it provides the best detection results in the most recently published studies (Dang et al. 2020; Rössler et al. 2019; Dolhansky et al. 2019). We followed the same training approach considered in Rössler et al. (2019): *(i)* the model was initialised with the weights obtained after training with the ImageNet dataset (Deng et al. 2009), *(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* (Nataraj et al. 2019b): 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 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* (Matern et al. 2019): 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 regularisation 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 Nataraj et al. (2019b) 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 colour features.

8.5.3 *Protocol*

The experimental protocol designed in this chapter aims at performing an exhaustive analysis of the state-of-the-art facial manipulation detection systems. As such, three different experiments were 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.

The GANprintR approach was trained during 100 epochs, using the Adam optimizer with a learning rate of 10^{-3}, and a mean square error (MSE) to obtain the reconstruction loss. To ensure an unbiased evaluation, GANprintR was trained with images from the MS-Celeb dataset (Guo et al. 2016), since it is disjoint from the datasets used in the development and evaluation of all the fake detection systems used in our experiments.

8.6 Experimental Results

This section describes the results achieved in the experimental framework of this chapter.

8.6.1 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 Sect. 8.2).

A total of six experiments were carried out: A.1 to A.6. Table 8.2 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. 8.4 the evolution of the loss/accuracy of the XceptionNet and Steganalysis detection systems for Exp. A.1.

The analysis of Fig. 8.4 shows that both XceptionNet and Steganalysis approaches were 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 accuracy achieved in Exp. A.1 for the XceptionNet and Steganalysis approaches were 99% and 95%, respectively. Similar trends were observed for the other experiments.

We now analyse the results included in Table 8.2 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 Sect. 8.2), 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 have poor accuracy. In the next section we will see the limitations of these techniques in-the-wild.

8.6.2 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 pre-

Table 8.2 Controlled and in-the-wild scenarios: manipulation detection performance in terms of EER (%) 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 (%)

Experiment	Development		Evaluation		XceptionNet (Chollet 2017)			Steganalysis (Nataraj et al. 2019b)			Local artifacts (Matern et al. 2019)		
	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

(continued)

513475_1_En_8_Chapter TYPESET DISK LE CP Disp.:13/11/2021 Pages: xxx Layout: T1-Standard

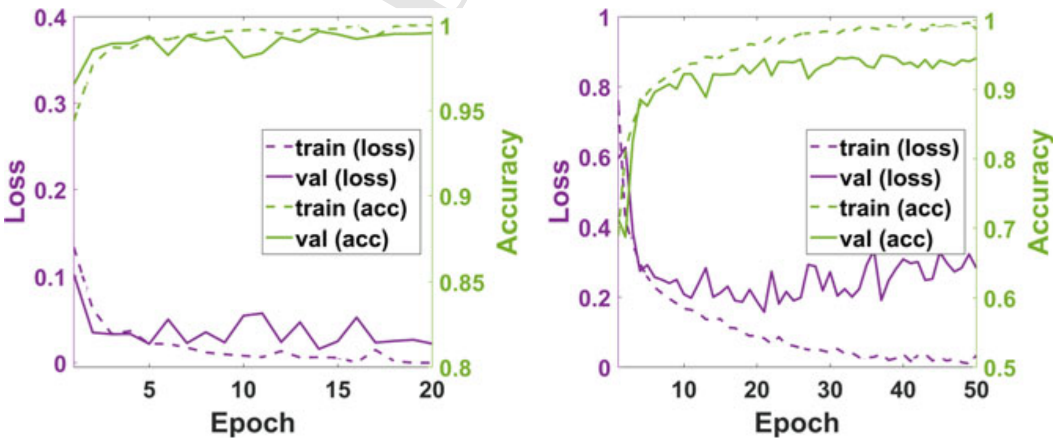
Table 8.2 (continued)

Experiment	Development		Evaluation		XceptionNet (Chollet 2017)			Steganalysis (Nataraj et al. 2019b)			Local artifacts (Matern et al. 2019)		
	Real	Synthetic	Real	Synthetic	EER	R_{real}	R_{fake}	EER	R_{real}	R_{fake}	EER	R_{real}	R_{fake}
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

liminary analysed in previous studies (Rössler et al. 2019; Korshunov and Marcel 2018), 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 8.2). In general, the results achieved in the experiments evidence a high degradation of the detection performance regardless of the 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 was 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 (Yu et al. 2018). Moreover, these results suggest that these GAN fingerprints are the information used by the detectors to distinguish between real and synthetic data.

Table 8.2 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, an average EERs of 9.3%, 32.3% and 42.3% in fake detection were 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 was found, which points that the change of synthetic data might be the main cause for performance degradation.



(a) XceptionNet [203] (b) Steganalysis [684]

Fig. 8.4 Exp. A.1: Evolution of the loss/accuracy with the number of epochs

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 8.3, together with the performance of the original images. The results suggest a high performance degradation in all experiments, proving the vulnerability of the fake detection system to unseen conditions, even if they result from simple image transformations.

To further understand the impact of these transformations, we evaluated an increasing downsize ratio in the performance of the fake detection system. Figure 8.5 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 by 1/4, the average EER 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.

8.6.3 GAN-Fingerprint Removal

This section analyses the results of the strategy for GAN-fingerprint Removal (GANprintR). We evaluated to what extent our method is capable of spoofing state-of-the-art facial manipulation detection systems 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 GANprintR.

Table 8.3 provides the results achieved for both the original fake data and after GANprintR. The analysis of the results shows that GANprintR obtains 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

Table 8.3 Comparison between the GANprintR approach and typical image manipulations. The detection performance is provided in terms of EER (%) for experiments A.1 to A.6, when using different versions of the evaluation set. TDE 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,

Experiment	TDE	EER (%)	R_{real} (%)	XceptionNet		
				R_{fake} (%)	PSNR (db)	SSIM
A.1	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	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	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	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	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	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

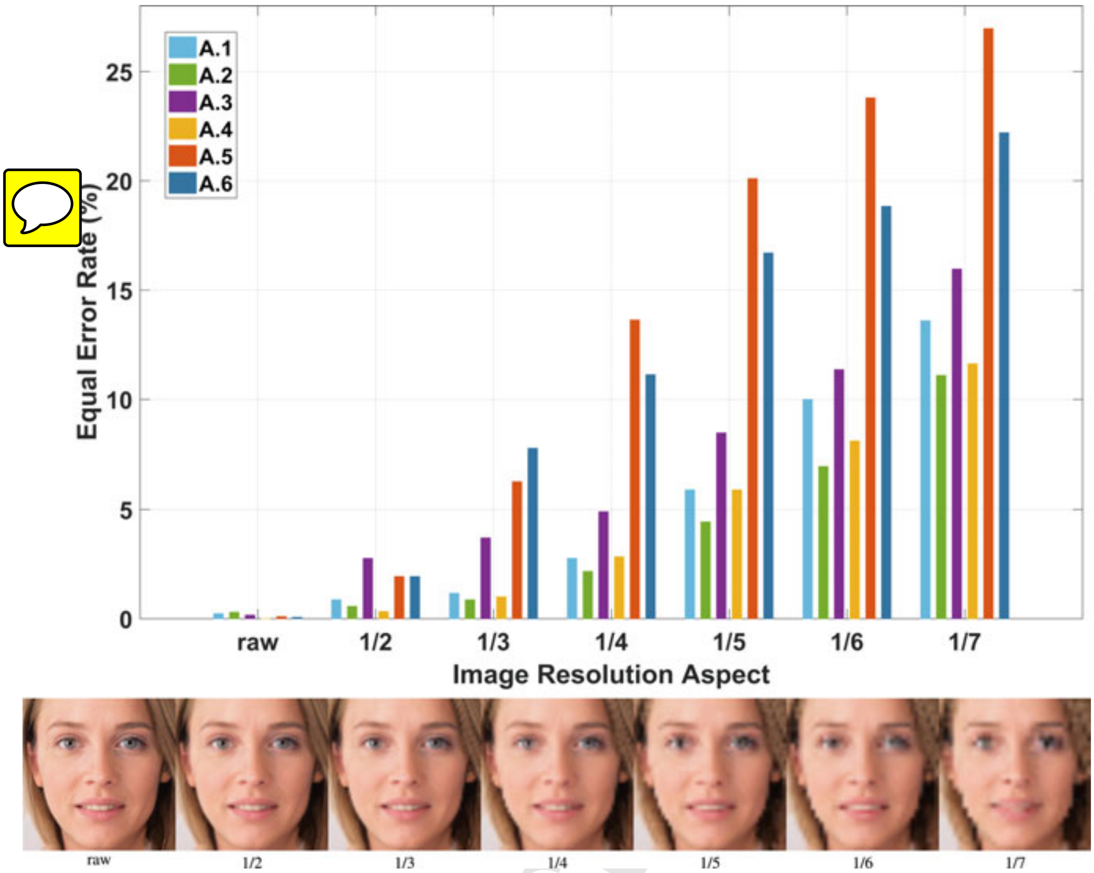


Fig. 8.5 Robustness of the fake detection system regarding the image resolution. The XceptionNet model is trained with the raw image resolution and evaluated with lower image resolutions. Note how the EER increases significantly while reducing the image resolution

their naturalness. It is important to remark that different real face databases were considered for training the face manipulation detectors and our GANprintR module.

In addition, we provide in Fig. 8.6 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. 8.6 together with a face example of each configuration to visualise the image quality. The face examples included in Fig. 8.6 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. The 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 detectors 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

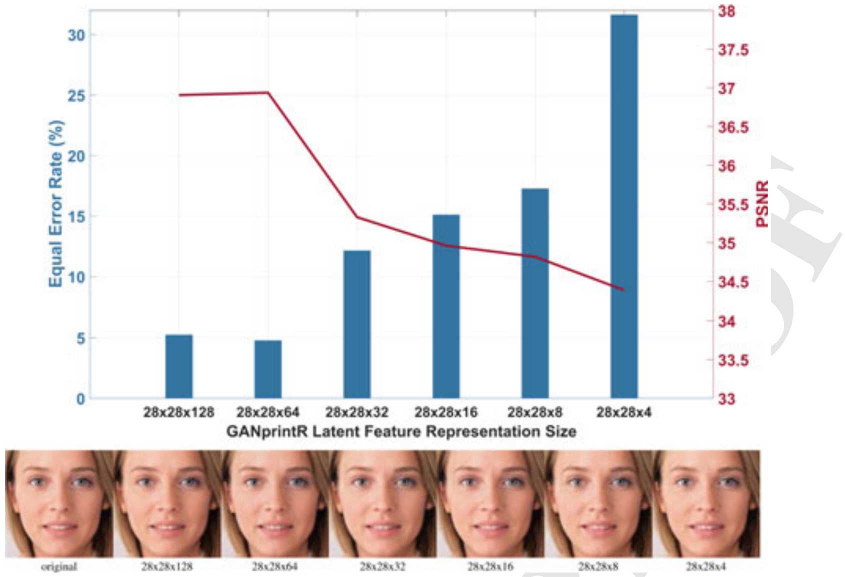


Fig. 8.6 Robustness of the fake detection system after GAN-fingerprint Removal (GANprintR). The latent feature representation size of the AE is varied to analyse the impact on both system performance and visual aspect of the reconstructed images. Note how the EER increases significantly when considering GANprintR spoof approach, while maintaining a high visual similarity with the original image

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. Figure 8.7 shows the fake detection performance in terms of EER for different sizes of the latent feature representation of GANprintR. Five different GANprintR 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 GANprintR is actually removing the GAN-fingerprint information.

8.6.4 Impact of GANprintR on Other Fake Detectors

For completeness, we provide in this section a comparative analysis between the impact of the GANprintR approach on the three state-of-the-art manipulation detection approaches considered in this chapter. Table 8.4 reports the EER and Recall observed when using the original images and when using the modified version of the same images.

In Sect. 8.6.1 it has been concluded that XceptionNet stands out as the most reliable approach at recognising synthetic faces. The analysis of Table 8.4 evidences that this conclusion also holds when using images transformed by GANprintR. Nevertheless, it is also interesting to analyse the performance degradation caused by the GANprintR approach. The average number of percentage points that the EER has increased for XceptionNet, Steganalysis and Local Artifacts is 9.65, 14.68 and 4.91, respectively.

513475_1_En_8_Chapter ☑ TYPESET ☐ DISK ☐ LE ☑ CP Disp.:13/11/2021 Pages: xxx Layout: T1-Standard

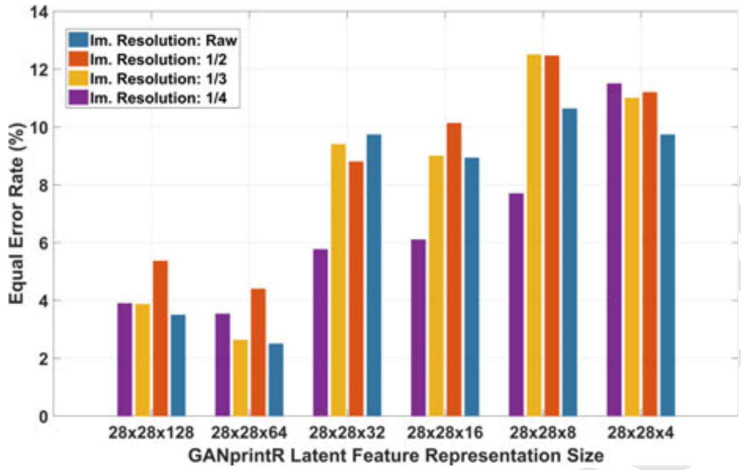


Fig. 8.7 Robustness of the fake detection system trained with different resolutions and then tested with fakes improved with GANprintR under various configurations (representation sizes). Five different GANprintR configurations are tested per image resolution level. The results observed point for the 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 fingerprints.

Even though, in this case, the work of Matern et al. (2019) stands out for having the lowest performance degradation, we believe that this is primarily due to the high EER achieved in the original set of images.

8.7 Conclusions and Outlook

This chapter has covered the topic of GAN fingerprints in face image synthesis. We have first provided an in-depth literature analysis of the most popular GAN synthesis architectures and fake detection techniques, highlighting the good fake detection results achieved by most approaches due to the "fingerprints" inserted in the GAN generation process.

In addition, we have reviewed a recent approach to improve the naturalness of facial fake images and spoof state-of-the-art fake detectors: GAN-fingerprint Removal (GANprintR). GANprintR was originally presented in Neves et al. (2020) and is based on a convolutional autoencoder. The autoencoder 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.

A thorough experimental assessment of this type of facial manipulation has been carried out considering fake detection (based on holistic deep networks, steganalysis, and local artifacts) and realistic GAN-generated fakes (with and without GANprintR) over different experimental conditions, i.e. controlled and in-the-wild scenarios. We

Table 8.4 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 GANprintR. The detection performance is provided in terms of EER (%), while R_{real} and R_{fake} denote the Recall of the real and fake classes, respectively

Experiment	Data	XceptionNet			Steganalysis (Nataraj et al. 2019b)			Local artifacts (Matern et al. 2019)		
		EER (%)	R_{real} (%)	R_{fake} (%)	EER (%)	R_{real} (%)	R_{fake} (%)	EER (%)	R_{real} (%)	R_{fake} (%)
A.1	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	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	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	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	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	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

513475_1_En_8_Chapter ☑ TYPESET ☐ DISK ☐ LE ☑ CP Disp.:13/11/2021 Pages: xxx Layout: T1-Standard

highlight 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 approaches such as GANprintR are 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, future work should research about strategies to harden such face manipulation detectors by exploiting databases such as iFakeFaceDBiFakeFaceDB.[4] Additionally, further works should study: *(i)* how improved fakes obtained in similar ways as GANprintR can jeopardise other kinds of sensitive data (e.g. other popular biometrics like fingerprint Tolosana et al. 2020a, iris Proença and Neves 2019, or behavioural traits Tolosana et al. 2020b), *(ii)* how to improve the security of systems dealing with other kinds of sensitive data (Hernandez-Ortega et al. 2021), and finally *(iii)* best ways to combine multiple manipulation detectors (Tolosana et al. 2021) in a proper way (Fiérrez et al. 2018) to deal with the growing sophistication of fakes.

Acknowledgements 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), Bio-Guard (Ayudas Fundación BBVA a Equipos de Investigación Científica 2017), by NOVA LINCS (UIDB/04516/2020) with the financial support of FCT—Fundação para a Ciência e a Tecnologia, through national funds, by FCT/MCTES through national funds and co-funded by EU under the project UIDB/EEA/50008/2020, and by FCT—Fundação para a Ciência e a Tecnologia through the research grant '2020.04588.BD'. 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.

[4] https://github.com/socialabubi/iFakeFaceDB.

References

Albright M, McCloskey S (2019) Source generator attribution via inversion. In: IEEE Conference on computer vision and pattern recognition workshops, CVPR workshops 2019, Long Beach, CA, USA, June 16–20, 2019. Computer Vision Foundation/IEEE, pp 96–103

Barni M, Kallas K, Nowroozi E, Tondi B (2020) CNN detection of GAN-generated face images based on cross-band co-occurrences analysis. arXiv:abs/2007.12909

Bellemare MG, Danihelka I, Dabney W, Mohamed S, Lakshminarayanan B, Hoyer S, Munos R (2017) The cramer distance as a solution to biased wasserstein gradients. arXiv:abs/1705.10743

Binkowski M, Sutherland DJ, Arbel M, Gretton A (2018) Demystifying MMD GANs. In: 6th international conference on learning representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, conference track proceedings. OpenReview.net

Bonettini N, Bestagini P, Milani S, Tubaro S (2020) On the use of benford's law to detect GAN-generated images. arXiv:abs/2004.07682

Brock A, Donahue J, Simonyan K (2019) Large scale GAN training for high fidelity natural image synthesis. In: 7th international conference on learning representations, ICLR 2019, New Orleans, LA, USA, May 6–9, 2019. OpenReview.net

Cellan-Jones R (2019) Deepfake videos double in nine months

Choi Y, Choi M-J, Kim M, Ha J-W, Kim S, Choo J (2018) StarGAN: unified generative adversarial networks for multi-domain image-to-image translation. In: 2018 IEEE conference on computer vision and pattern recognition, CVPR 2018, Salt Lake City, UT, USA, June 18–22, 2018. IEEE Computer Society, pp 8789–8797

Chollet F (2017) Xception: deep learning with depthwise separable convolutions. In: 2017 IEEE conference on computer vision and pattern recognition, CVPR 2017, Honolulu, HI, USA, July 21–26, 2017. IEEE Computer Society, pp 1800–1807

Cozzolino D, Gragnaniello D, Verdoliva L (2014) Image forgery detection through residual-based local descriptors and block-matching. In: 2014 IEEE international conference on image processing, ICIP 2014, Paris, France, October 27–30, 2014. IEEE, pp 5297–5301

Dang H, Liu F, Stehouwer J, Liu X, Jain AK (2020) On the detection of digital face manipulation. In: 2020 IEEE/CVF conference on computer vision and pattern recognition, CVPR 2020, Seattle, WA, USA, June 13–19, 2020. IEEE, pp 5780–5789

Deng J, Dong W, Socher R, Li L-J, Li K, Li F-F (2009) Imagenet: a large-scale hierarchical image database. In: 2009 IEEE computer society conference on computer vision and pattern recognition (CVPR 2009), 20–25 June 2009, Miami, Florida, USA. IEEE Computer Society, pp 248–255

Dolhansky B, Howes R, Pflaum B, Baram N, Canton-Ferrer C (2019) The deepfake detection challenge (DFDC) preview dataset. arXiv:abs/1910.08854

Durall R, Keuper M, Keuper J (2020) Watch your up-convolution: CNN based generative deep neural networks are failing to reproduce spectral distributions. In: 2020 IEEE/CVF conference on computer vision and pattern recognition, CVPR 2020, Seattle, WA, USA, June 13–19, 2020. IEEE, pp 7887–7896

FaceApp (2017)

Fiérrez J, Morales A, Vera-Rodríguez R, Camacho D (2018) Multiple classifiers in biometrics. Part 2: trends and challenges. Inf Fusion 44:103–112

Flickr-Faces-HQ Dataset (FFHQ) (2019)

Frank J, Eisenhofer T, Schönherr L, Fischer A, Kolossa D, Holz T (2020) Leveraging frequency analysis for deep fake image recognition. In: Proceedings of the 37th international conference on machine learning, ICML 2020, 13–18 July 2020, Virtual Event, volume 119 of Proceedings of machine learning research. PMLR, pp 3247–3258

Galbally J, Marcel S, Fierrez J (2014) Biometric anti-spoofing methods: a survey in face recognition. IEEE Access 2:1530–1552

Gandhi A, Jain S (2020) Adversarial perturbations fool deepfake detectors. In: 2020 international joint conference on neural networks, IJCNN 2020, Glasgow, United Kingdom, July 19–24, 2020. IEEE, pp 1–8