Chapter 9
Detection of AI-Generated Synthetic Faces

Diego Gragnaniello, Francesco Marra, and Luisa Verdoliva

Abstract In recent years there have been astonishing advances in AI-based synthetic media generation. Thanks to deep learning methods it is now possible to generate visual data with a high level of realism. This is especially true for human faces. Advanced deep learning tools allow one to easily change some specific attributes of a real face or even create brand new identities. Although this opens up a large number of new opportunities, just think of the entertainment industry, it also undermines the trustworthiness of media content and supports the spread of fake identities over the internet. In this context, there is a fundamental need to develop robust and automatic tools capable of distinguishing synthetic faces from real ones. The scientific community is making a huge research effort in this field, proposing several interesting approaches. However, a universal detector is yet to come. Fundamentally, the research in this field is like a cat and mouse game, with new detectors that are designed to deal with powerful synthetic face generators, while the latter keep improving to produce more and more realistic images. In this chapter we will present the most effective techniques proposed in the literature for the detection of synthetic faces. We will analyze their rationale, present real-world application scenarios , and compare different approaches in terms of accuracy and generalization ability.

9.1 Introduction

Among the many applications of generative adversarial networks (GANs), image synthesis is one of the most investigated, and research in this field has shown a great potential. Particularly impressive are the results that can be achieved in face

D. Gragnaniello · F. Marra · L. Verdoliva (✉)
University of Naples Federico II, via Claudio 21, Naples, Italy
e-mail: verdoliv@unina.it

D. Gragnaniello
e-mail: diego.gragnaniello@unina.it

F. Marra
e-mail: francesco.marra@unina.it

C. Rathgeb et al. (eds.), *Handbook of Digital Face Manipulation and Detection*,
Advances in Computer Vision and Pattern Recognition,
https://doi.org/10.1007/978-3-030-87664-7_9

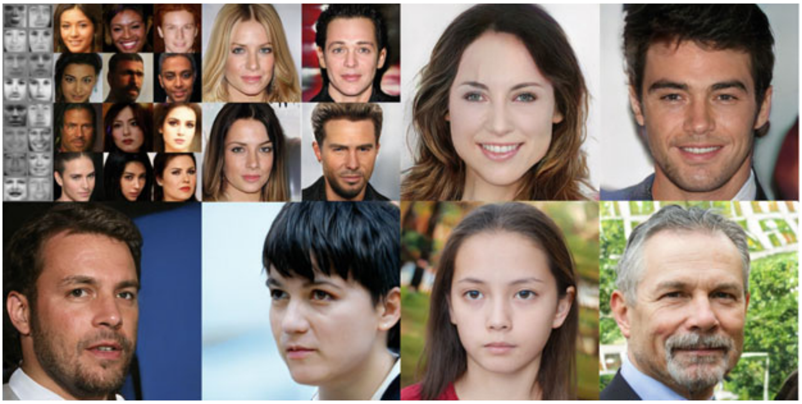


Fig. 9.1 Fully synthetic face images generated by different GAN architectures. Top, from left to right: images generated using the method proposed in [20], BEGAN [3], and ProGAN [25] at two different resolutions. Bottom, images generated by StyleGAN [27] (left) and StyleGAN2 [28] (right)

generation, with images of higher and higher resolution quality, as also shown by the examples of Fig. 9.1 which depict the evolution of synthetic faces over time. The visual appearance of the images generated by the latest GAN architectures is so realistic that it deceives even the experienced and attentive observer. This raises major concerns on the possible malicious use of such tools. For example, they can be used to create fake profiles on social networks and, more in general, they can be used to spread false information over the web. Therefore, it is urgent to develop automatic tools that can reliably distinguish real content from synthetic content.

Despite their high visual quality, GAN images are characterized by specific artifacts left from the generation process that can be used to develop effective tools for their detection. In some cases, their synthetic origin can be identified by visual inspection due to the presence of semantic inconsistencies, such as color anomalies or lack of symmetries. More generally, these images present invisible artifacts, closely linked to the architecture of the generative network, which can be extracted through appropriate processing steps. These artifacts represent very strong clues, which can be exploited even when synthetic images appear perfectly realistic. In fact, GAN-generated images have been shown to embed a sort of artificial fingerprints [36, 60], specific to each individual GAN architecture. Such patterns also show themselves as peaks in the Fourier domain, not present in the spectral distribution of natural images [16, 18, 61] (see Fig. 9.2).

Many of the detectors proposed so far for GAN-generated faces explicitly use the features described above, while others exploit them implicitly by relying on convolutional neural networks suitably trained on very large datasets [52]. Typically, these solutions show very good performance in distinguishing synthetic faces from real ones. However, they often require that the training set include a sufficient number of examples of the specific GAN architecture that generated images in the test set.

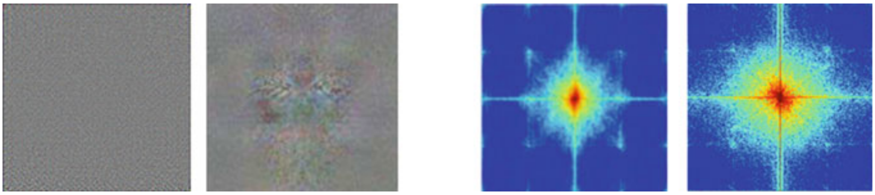


Fig. 9.2 GAN fingerprints extracted in the spatial domain (left) and traces of synthetic images in the frequency domain (right)

Hence, the limited generalization capability is a major problem for current GAN image detectors. As new AI-based models for synthesizing faces are proposed by the day, it is very important to propose solutions that can generalize to new unseen examples. Likewise, robustness is a major challenge, as images are routinely compressed and resized on social networks and valuable clues can be easily reduced or destroyed.

In this chapter, after briefly reviewing the main GAN architectures for face generation, we carry out an analysis of the state-of-the-art detection techniques. We will first present the notion of artificial fingerprints and then describe the major detection methods. We will also present an investigation on the performance of the most promising detectors by testing their generalization and robustness ability on several recent GAN architectures. Besides providing a baseline, this comparative analysis allows us to single out some key features of successful solutions, clearing the way for the design of new and more effective tools.

9.2 AI Face Generation

Progress on synthetic face generation has been possible thanks to the development of deep learning techniques especially autoencoders and generative adversarial networks [20], but also the availability of large-scale public face datasets. Early works were trained on very small face images dataset, while more recent ones rely on the CelebA dataset [33], that includes more than 200k face images of 10k identities, its extension CelebA-HQ with 30k images, and FFHQ [27] that comprises 70k high-quality images collected from Flickr.

AI face generation methods can be roughly classified in the following categories:

- *Fully synthetic faces*: generated faces are synthesized completely from scratch. Some examples have been already shown in Fig. 9.1. Beyond the availability of high resolution face images, some specific strategies have been of key importance to produce more accurate and realistic faces than those produced by the basic GAN architecture [20]. A major breakthrough came with the ProGAN architecture proposed in [25], where high resolution has been achieved by growing both the generator and discriminator progressively during the training process. Another

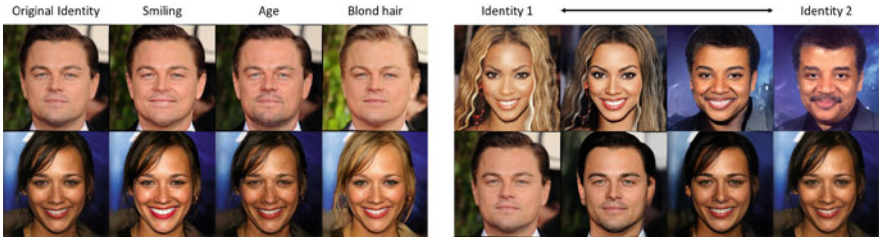


Fig. 9.3 Images manipulated by changing a face attribute (left) and images where two identities are fused together (right)

significant improvement can be found in several works that rely on style transfer to gain more control in the synthesis process and that led to several successful architectures: StyleGAN[1] [27], StyleGAN2 [28] and the recent variant adaptive discriminator augmentation (ADA) [26].

- *Face attributes modification*: beyond synthesizing faces from scratch, it is also possible to modify an attribute of a real face, such as gender, age, skin, or hair color. Conditional GANs represent a very effective tool to address this task and many different approaches have been proposed in the literature and that allow a surprisingly realistic result [32, 46, 51, 55, 62]. More sophisticated modifications let to change the pose or the facial expression [49, 58]. In Fig. 9.3 (left), some examples are shown. It is worth underlining that these manipulations do not change the original identity of the involved subject. Some of these approaches can be found in some mobile applications, such as the popular FaceApp[2].
- *Face blending*: this category comprises methods that are able to fuse the identities from two different face images. The resulting identity is neither non-existent nor preserved, but the resulting face mixes both identities in one. In Fig. 9.3 (right) some examples of face identity blending[3] are presented using the approach proposed in [30].

9.3 GAN Fingerprints

Early work on synthetic media forensics has focused on extending successful approaches and methods of real multimedia forensics to this new domain. In particular, device and model fingerprints represent formidable assets to perform a wide array of forensic tasks, from source attribution to forgery detection and localization, to blind image clustering. Device fingerprints have been first exposed in the seminal work of Chen et al. [34] and Lukas [9]. Due to sensor imperfections, each camera

[1] https://thispersondoesnotexist.com/.

[2] https://play.google.com/store/apps/details?id=io.faceapp.

[3] https://openai.com/blog/glow/.

presents a so-called photo-response non-uniformity (PRNU) which leaves on each acquired image traces that are unique of that device and stable in time. This image-like pattern represents therefore a device fingerprint, which can be reliably estimated from sample images of the device.

Given their potential, extending such tools to synthetic media has an obvious appeal. The existence of "artificial" GAN fingerprints was first demonstrated in [36]. These fingerprints are extracted using the very same procedure adopted for real fingerprints. More specifically, for a generic image X_i generated by a given GAN a high-pass filter, i.e., a denoiser, is used to remove the semantic image content:

$$R_i = X_i - f(X_i) \tag{9.1}$$

Then, we assume the residual to be the sum of a *non-zero* deterministic component, the fingerprint F, and a random noise component W_i

$$R_i = F + W_i \tag{9.2}$$

Accordingly, the fingerprint is estimated by a simple average over the available residuals

$$\widehat{F} = \frac{1}{N}\sum_{i=1}^{N} R_i \tag{9.3}$$

As the number of averaged residuals grows, a weak but stable pattern emerges, which characterizes uniquely the GAN architecture. The whole procedure is outlined in Fig. 9.4. Once the GAN fingerprint has been extracted from 200 to 300 GAN images, it can be compared by means of the normalized cross-correlation with the noise residual extracted from the image under test. Experiments carried out in [36] prove

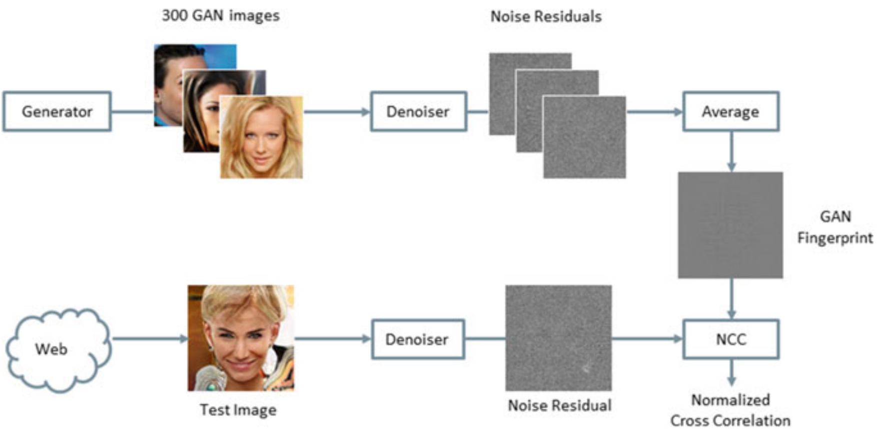


Fig. 9.4 Pipeline for GAN fingerprint extraction

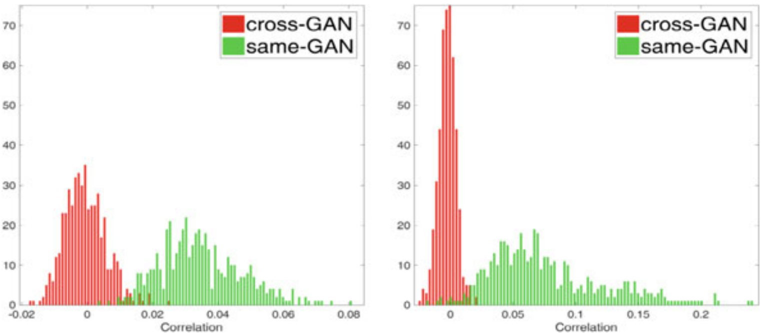


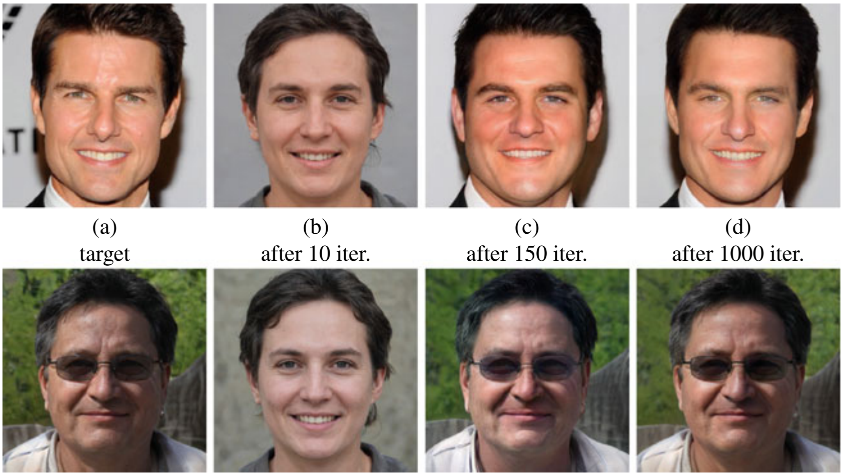
Fig. 9.5 Correlation of CycleGAN (left) and ProGAN (right) residuals with same/cross-GAN fingerprints

that such fingerprints can be used to reliably tell apart real images from synthetically generated ones, and also to attribute an image to its source GAN.

As an example, Fig. 9.5 shows the histograms of the correlation coefficients between image residuals and fingerprint of two GAN architectures. On the left, the GAN-A fingerprint is considered, with green/red colors indicating images generated from the same (GAN-A) or the other (GAN-B) network. The cross-GAN histogram is evenly distributed around zero, indicating no correlation between generated images and unrelated fingerprints. On the contrary, the same-GAN histogram is shifted around larger values, testifying of a significant correlation with the correct fingerprint. The behavior is very similar when GAN-B residuals are considered and the roles are reversed, on the right. In both cases the two distributions are well separated, allowing reliable discrimination.

In [60] fingerprint extraction is addressed by means of a supervised deep learning scheme, where the fingerprint maximizes the correlation with images generated by the same-GAN. Under this setting, both image-like fingerprints, like in [36], and compact vectorial fingerprints can be used. The sophisticated extraction process further improves the performance. Moreover, the experiments prove that different fingerprints arise not only due to different GAN architectures but also from small differences in the training of the same architecture, enabling fine-grained model authentication. Also, GAN fingerprints are shown to persist across different image frequencies and patches and are not biased by GAN artifacts. Both [36] and [60] suggest that the regular patterns observed in GAN fingerprints are due to the up-sampling operations typical of the synthesis network, while instance-level peculiarities depend on the specific filters learned in training.

In [1, 28] attribution of GAN generated images to their source is pursued through GAN inversion. The idea is to provide the test image as target to a set of generators. The likely source is the generator that ensures the minimum reconstruction error. In fact, a GAN architecture cannot perfectly generate a synthetic image that has been produced by another GAN architecture nor it can perfectly reproduce a real image. The projection-based method of [28] was used to prove that an image was synthesized



(a) target | (b) after 10 iter. | (c) after 150 iter. | (d) after 1000 iter.

Fig. 9.6 Target face (**a**) and generated faces at different iterations (**b**, **c**, **d**). In one case (top) the GAN model is not able to perfectly reproduce the target real face, while it succeeds in perfect reconstruction (bottom) with a target image generated by the GAN itself, that is, face (**a**) is identical to (**d**)

by a specific GAN network. We show such a result in Fig. 9.6, where the target image (a) and the output of the GAN generation process at different iterations are shown. We can observe that in one case (top figure) the GAN is not able to perfectly reproduce the target face, since it is real, while in the second case the target face is perfectly reproduced by the GAN generator (bottom figure), which demonstrates that it was generated by that GAN model.

9.4 Detection Methods in the Spatial Domain

Most of the techniques that aim at distinguishing AI-generated faces from real ones rely on some sort of artifacts, either visible, such as unnatural facial traits, or invisible, like pixel-level statistical inconsistencies that suggest the presence of a generative process. In this section we present detection approaches that work in the original spatial domain. They all use a neural classifier, eventually, but differ for the nature of the features on which the classification is based, handcrafted, or data-driven.

9.4.1 Handcrafted Features

Several handcrafted discriminative features have been proposed to detect generated face images, typically based on the visual inspection of GAN imagery and on prior knowledge of the relevant architectures. In the following we describe the most common and effective ones.

- *Face asymmetries*. Synthetic faces are often characterized by unnatural asymmetries. Indeed, to the best of our knowledge, no specific constraint on symmetry is imposed in the generation phase, probably because of technical difficulties. Therefore, symmetry emerges only as a common feature of the training data and cannot be ensured for all tiny details, however significant for a human observer. For example, GAN images sometimes present eyes with different colors, or asymmetric specular reflections, different earrings, or only on earrings, or ears with markedly different characteristics (see Fig. 9.7). These artifacts are exploited in [39], where simple features are built in order to capture them, such as the correlation between the eyes in suitable color spaces. To exploit asymmetric corneal specular reflections a detector is proposed in [23] based on inconsistencies between light sources reflected in the two eyes. However, this approach needs high-resolution images in order to correctly segment the light spots in both eyes and then compare them, which is not the case of most social networks. This problem is tackled in [22], where a super-resolution module is used, trained to preserve generation artifacts. After the resolution increase, a CNN is used which pools different feature maps on the basis of facial key-points.
- *Landmark locations*. Just like for symmetry, no explicit constraint can be imposed in the generation process to ensure the correct positioning of facial landmark points. As a consequence, it may happen that all individual face parts are generated with a high level of realism and with many details, but their relative locations are unnatural. Based on this observation, the method proposed in [57] uses the locations of the facial landmark points, like the tips of the eyes, nose, and the mouth, as discriminative features for detection.
- *Color features*. GANs produce by design only a limited range of intensity values, and do not generate saturated and/or under-exposed regions. While this is a good property to ask of a photo, a large number of natural face images do present extreme-valued pixels, and their absence suggests a synthetic origin. This fact is exploited in [40] by measuring the frequency of saturated and under-exposed pixels in each image. Turning to color, current GANs are known to not accurately preserve the natural correlation among color bands. This property is exploited in [31] where the chrominance components of the image are high-pass filtered and their co-occurrence matrices are computed to form discriminative features for detection. Indeed, co-occurrences of high-pass filtered images are popular tools in image forensics since invisible artifacts are often present in the high-frequency signal components [12]. Thus, co-occurrence matrices extracted from the RGB channels are also used in [42] as the input of a CNN and, similarly, in [2] co-occurrences across color bands are computed to capture discriminative information.

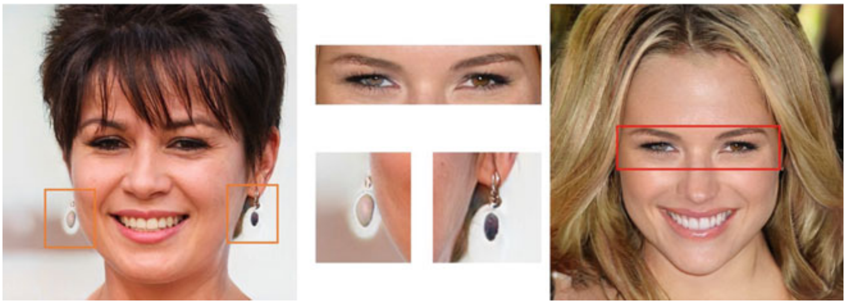


Fig. 9.7 Examples of GAN synthetic faces with visible artifacts. A generated face with asymmetric earrings (left) and a face with eyes of different colors

9.4.2 *Data-Driven Features*

Deep networks, in particular Convolutional Neural Networks (CNNs), have proven to adapt well to multimedia forensic tasks [48]. A first investigation of detectors based on very deep networks is carried out in [35], where state-of-the-art pre-trained CNNs, like Xception, Inception, and DenseNet, are shown to ensure excellent performance for GAN image detection. In particular, they turn out to outperform CNN models specifically tailored to forensics tasks and trained from scratch, especially in the most challenging scenarios. More recently, Xception [11] has been used also in [15] as the backbone of a strategy that includes an attention mechanism.

In [53], following an approach originally proposed in Deepxplore [45], detection is based on the neurons' activity at each layer of the network. Experiments carried out on the challenging DFDC dataset show that the neurons' activity provides detailed information about the network behavior and leads to improved classification performance and higher robustness against adversarial attacks. In [24] both detection and attribution are pursued by means of a three-level hierarchical framework. The first level distinguishes real images from manipulated ones, the latter are then classified in the second level as retouched or generated from scratch, and these latter are finally attributed to the generating GAN architecture in the third level. At each level, a CNN is used for feature extraction and an SVM for classification.

GAN architectures typically includes up-sampling stages, which produce a typical checkerboard pattern. To exploit this trace, in [41] an ad hoc self-attention mechanism is proposed to replace plain global pooling in the final layers of the CNN.

9.5 Detection Methods in the Frequency Domain

The checkerboard pattern mentioned in the previous Section shows its traces very clearly in the frequency domain. In fact, the up-sampling operations give rise to quasi-periodic patterns which result in strong peaks in the image spectrum

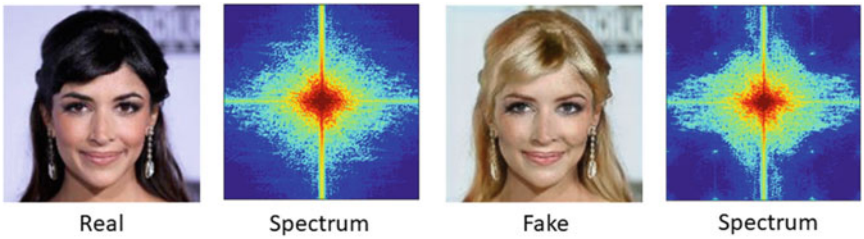


Fig. 9.8 A real image and its Fourier transform (left) a GAN image generated using starGAN [10] and its Fourier transform (right). In this last case it is possible to observe clear peaks in the spectrum

(see again Fig. 9.2 for generic images and Fig. 9.8 for faces). Based on this observation, a detector is proposed in [61] which takes the frequency spectrum instead of image pixels as input for a CNN. A frequency-domain analysis is also performed in [18] to investigate the presence of artifacts across different network architectures, datasets, and resolutions. Then, a CNN-based classifier is trained with Fourier spectra taken from both real images and their synthetic versions obtained through an adversarial autoencoder. Also [17] shows that GAN images do not faithfully mimic the spectral distributions of natural images. Various generative architectures are considered, based both on GANs and autoencoders, and the spectra of the generated images are compared with those of real ones. It results that the spectrum decay along the radial dimension is markedly different in the two cases, with fake images that exhibit higher energy at mid-high frequencies than real ones, which corresponds to small-scale correlations. To exploit these findings, a KNN classifier is trained using the energy spectral distribution as an input feature. Along the same line, in [16] a parametric model is used to fit the decay function of the Fourier spectrum and a classifier is trained on the fitting parameters. It is worth noting that both approaches propose also countermeasures to limit the appearance of such spectral artifacts by means of a simple post-processing [17] or a spectral loss to be used during GAN training [16].

Frequency analyses have been also widely used to detect generated images shared online. Indeed, images uploaded to the web are very often coded using the JPEG standard, based on the Fourier-like discrete cosine transform (DCT). For synthetic images, this compression step may reveal distinctive traces of the generation process, absent in real images, which can be used for reliable detection. As an example, for generated images, the most significant digit of the quantized DCT coefficients violates the well-known Benford's law. Based on this evidence, in [4] a compact feature vector is extracted from the DCT coefficients and used to train a random forest classifier. Frequency-aware features are learned in the DCT domain in [47] to exploit both local and global frequency clues. On one hand, the proposed approach learns the global DCT coefficients where it is easier to spot fake faces. On the other hand, block-wise DCT frequency statistics are computed as complementary features to improve detection.

9.6 Learning Features that Generalize

Fully supervised approaches are typically very effective when the GAN images under test come from a model that is also present in training. However, often they fail to generalize to data generated by new unseen models. This phenomenon has been shown both in [29] and in [14], where some interesting experiments are carried out that highlight the inability of both handcrafted and data-driven features to support cross-dataset generalization. In the following we will review some of the methods proposed so far to address this issue.

- *Few-shot and incremental learning*. In [14] a strategy based on few-shot learning is proposed to increase transferability. An autoencoder with a bipartite hidden layer is trained. Then, the input image is projected onto a latent vector where the information needed to make the real/synthetic decision is disentangled from the image representation. This allows for higher detection rates in cases where only a few training samples of an unseen GAN architecture are available. In [38], instead, an approach based on incremental learning is proposed to update the detector to new data (i.e., new GAN architectures) made available at different times. A few representative template vectors of the known architectures are kept in a compact memory. In this way, the network can be re-trained on new data of a novel architecture without forgetting the old ones. Despite the improved generalization, these methods still require some examples of the new GAN architecture, which could not be available in a real scenario.
- *Augmentation*. A different solution is proposed in [56]. The idea is to carry out augmentation by Gaussian blurring so as to force the discriminator to learn more general features while discarding noise-like patterns that impair the training. A similar approach is followed in [54] where a standard pre-trained model, ResNet50, is further trained with a strong augmentation based on compression and blurring. Experiments show that, even by training on a single GAN architecture, the learned features generalize well to unseen architectures, datasets, and training methods. The comparative analysis of [21], instead, shows that by avoiding any subsampling in the first layer of the network ensures improved detection results. This finding is also confirmed by studies on no-subsampling network architectures for more general multimedia forensics tasks [37].
- *Patch-based learning*. A different perspective is adopted in [8] where a fully convolutional patch-based classifier with limited a receptive field is proposed. The authors prove the importance of focusing on local patches rather than on the global structure of the image, and hence ensemble the patch-wise decisions to obtain the overall prediction.

9.7 Generalization Analysis

Early techniques proposed for the detection of AI-generated faces were evaluated in an ideal scenario in which both the training and testing samples were generated by the very same AI (or small variations thereof). In this setting, even a simple approach like a shallow CNN can reach almost perfect performance [2, 4, 18, 35, 42]. As already discussed in the previous section, the detection performance drops on images generated by different GAN architectures. In this chapter, we will analyze the ability of several AI face detectors to generalize on synthetic images that are not used during training.

Following the protocol proposed in [54], we train all the detection methods on a large dataset of pristine images from LSUN, while synthetic images are generated using 20 ProGAN models [25], each trained on a different category, for a grand total of more than 700k images. All images have a resolution of 256 × 256 pixel and a subset of 4k images are used for validation. The test dataset comprises both same-resolution and higher resolution (1024 × 1024) images generated by various GAN architectures: StyleGAN [27], StyleGAN2 [28], BigGAN [6], CycleGAN [62], StarGAN [10], RelGAN [55], and GauGAN [44]. Then we have a large dataset of real images both low-resolution and high-resolution ones, as specified in [21].

In this analysis, the following synthetic image detectors are considered: Xception [35], SRNet [5], Spec [61], M-Gb [56], Co-Net [42], Wang2020 [54], PatchForensics [8]. Beyond these methods that are specifically proposed for GAN image detection, we also include SRNet that was instead originally proposed for steganalysis. In fact, both steganalysis and image forensics have a very similar goal, i.e., detecting hidden traces in the image, and methods proposed for steganalysis have often shown a great potential also in forensics [52]. More specifically, to better preserve features related to noise residual, SRNet avoids down-sampling in the first layers of the network.

To manage both low- and high-resolution images in the test phase, we adopt the strategy proposed in the original papers. In particular, for M-Gb, FFD and PatchForensics, the image is resized to the dimension of network input, meanwhile for Spec the central clip of size 224 × 224 is considered. The remaining techniques are applied on the whole test image without clipping/resizing it since they include a global average pooling. The list of the analyzed approaches and their test strategy are summarized in Table 9.1.

Results are shown in Fig. 9.9 for low-resolution (top) and high-resolution (bottom) images in terms of several performance metrics: area under the receiver-operating curve (AUC), accuracy at the fixed threshold of 0.5, and probability of detection for a 5% (Pd@5%) and 1% (Pd@1%) false alarm rate (FAR). Performance in terms of AUC on low-resolution (LR) images are very good, considering that there is a misalignment between training and testing data, with several methods exceeding the 0.9 level. However, accuracy results are much less encouraging, since a fixed threshold is used. Indeed, we noticed that each GAN architecture needs a different threshold to be set. Hence, without sample images generated from a specific GAN, it is hard to set the correct threshold. Considering the Pd@FAR metric, results become

Table 9.1 List of the methods used in our analysis together with the test strategy, as proposed in the original papers

References	Acronym	Test strategy
[35]	Xception	No cropping and no resizing
[5]	SRNet	No cropping and no resizing
[61]	Spec	Central cropping (224 × 224)
[56]	M-Gb	Resizing (128 × 128)
[42]	Co-Net	No cropping and no resizing
[15]	FFD	Resizing (299 × 299)
[54]	Wang2020	No cropping and no resizing
[8]	PatchForensics	Resizing (299 × 299)

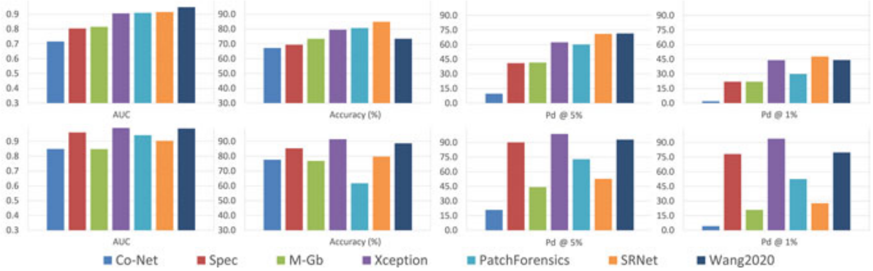


Fig. 9.9 Results of the methods under comparison in terms of AUC, Accuracy, Pd@5% and Pd@1% for all the tested methods on low-resolution (top) and high-resolution images (bottom)

worse, and only a few methods are able to ensure a good detection ability for high-resolution images. It is interesting to observe that the ranking of the methods change based on the specific metric.

9.8 Robustness Analysis

In this section we present a robustness analysis of the GAN detectors analyzed in the previous section. In fact, it is important to understand to which extent these detectors are affected by post-processing operations such as image compression or resizing that is commonly applied when images are uploaded on a social network. These operations could strongly reduce the low-level inconsistencies. For example in Fig. 9.10 it is shown the spectrum of GAN images when resizing and compression operations have been applied. One can observe that by reducing the size of the image the peaks in the Fourier domain tend to vanish, while enlarging the image further enhances those artifacts. Compression reduces the Fourier artifacts that completely disappear if the quality factor is too low (below 70).

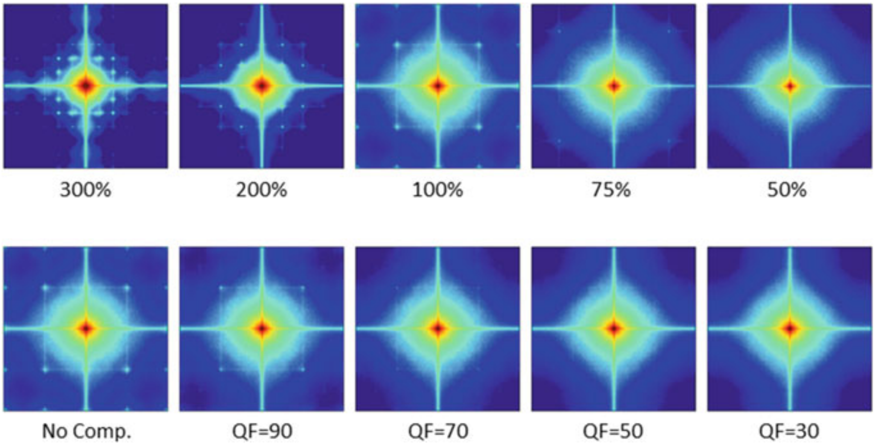


Fig. 9.10 Fourier transform of a GAN image by varying its dimensions using different resizing factors (top) and by applying JPEG compression at different quality levels (bottom)

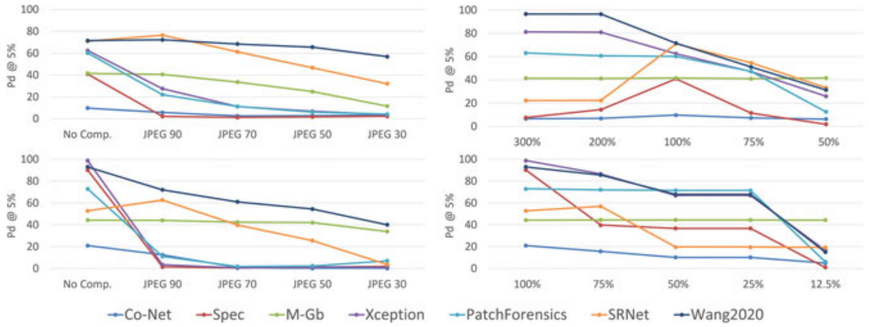


Fig. 9.11 Results of the methods under comparison in terms of Pd@5% by varying the JPEG quality compression level and by resizing the images at different factors. LR images are both enlarged and reduced in size, while HR images are only reduced

Figure 9.11 reports the Pd@5% performance for low-resolution and high-resolution images for varying compression factors and resizing scales. Several methods suffer dramatic impairments as soon as they move away from the ideal case of no compression and 100% scale. For example, we can notice that a 2x downsampling has a catastrophic effect, as justified by the fact that peaks completely disappear in the Fourier spectrum (see again Fig. 9.10). The most robust methods are those that benefit by a strong augmentation, in addition we can observe the good performance of SRNet on compressed images. Overall these experiments suggest that there is still much room for improvements with respect to the existing solutions, especially in terms of robustness to compression and resizing.

9.9 Further Analyses on GAN Detection

In this section we want to further investigate the performance of a good solution for GAN detection so as to identify the key ingredients of the most promising solutions. We consider as baseline the method proposed in [54], given the very good performance shown in the previous experiments, and introduce the following variations: remove Imagenet pre-training (no-pretrain), include an initial layer for residual extraction as often performed in image forensics strategies [52] (residual), do not perform down-sampling in the first layer as suggested by [5] (no-down), perform a stronger augmentation (strong-aug) by including Gaussian noise adding, geometric transformations, cut-out, and brightness and contrast changes. In addition, for the no-down variant, we also change the backbone network and replace ResNet50 with Xception (Xception no-down) and Efficient-B4 (Efficient no-down).

Results for the various metrics are shown in Fig. 9.12, while Fig. 9.13 shows results in terms of Pd@5% as a function of compression level and scaling factor. We can notice that the solution that avoids down-sampling in the first block of the architecture is very promising also in presence of resizing and compression. Instead no significant improvement can be observed by adopting strong augmentation or changing the backbone network. Note also the importance of the pre-training step on imagenet especially to gain robustness to resizing and compression.

Finally, in Table 9.2 we show the results for the baseline and the best variant over all the different GAN architectures, also including ProGAN that was used in the training step. We can notice that the best variant (no-down) provides an average gain of about 15% in terms of accuracy and 14% in terms of Pd@5%. Overall accuracy is always above 90% irrespective of the type of architecture. Finally, we added a further experiment by adopting 23 StyleGAN2 different models in training. In this last case performance are almost perfect with a further consistent improvement with respect to our baseline.

These experiments confirm the importance of diversity to increase robustness, like ImageNet pre-training, as already observed in steganalysis [59]. For the same

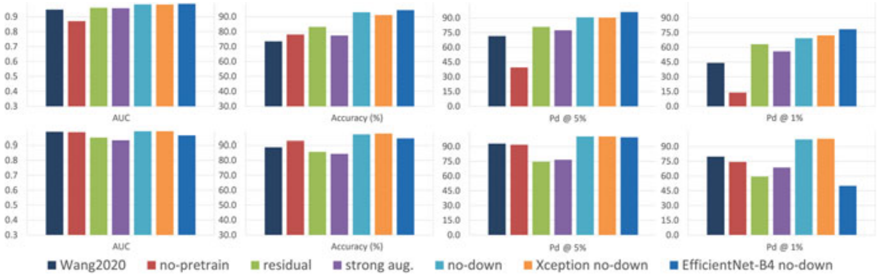


Fig. 9.12 Results of the baseline (Wang2020) and its variants in terms of AUC, Accuracy, Pd@5% and Pd@1% for variants of Wang2020 on low-resolution (top) and high-resolution images (bottom)

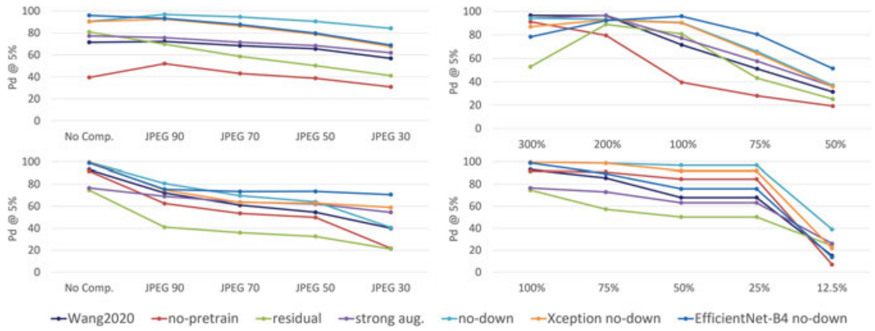


Fig. 9.13 Results of the baseline (Wang2020) and its variants in terms of Pd@5% by varying the JPEG quality compression level and by resizing the images at different factors. LR images are both enlarged and reduced in size, while HR images are only reduced

Table 9.2 Accuracy and Pd@5% for the baseline and the best variant that avoids down-sampling in the first block

Accuracy/Pd@5%		Wang2020 (baseline)	Best variant (no-down)	On StyleGAN2 (no-down)
Low res.	ProGAN	99.3/100.0	94.7/100.0	99.8/100.0
	StyleGAN	75.9/73.9	93.7/93.1	99.9/100.0
	StyleGAN2	71.5/69.0	92.2/88.8	99.9/100.0
	BigGAN	59.2/45.2	93.5/92.0	96.5/99.4
	CycleGAN	77.4/80.5	90.3/81.5	96.5/99.5
	StarGAN	84.3/89.4	94.5/97.6	99.9/100.0
	RelGAN	63.6/56.0	92.8/86.6	99.7/100.0
	GauGAN	82.5/86.3	93.6/93.5	90.8/97.1
High res.	ProGAN	99.7/100.0	97.1/100.0	99.7/100.0
	StyleGAN(Cel.)	99.3/100.0	97.1/100.0	99.7/100.0
	StyleGAN(FFHQ)	82.6/93.7	96.6/98.7	99.7/100.0
	StyleGAN2	73.2/78.1	96.9/99.6	99.7/100.0

reason, image pre-processing like resizing to match the input size of the CNN should be avoided. In fact, just like other forensics applications, the useful information lays in pixel-level patterns spread all over the image. If size reduction is necessary, cropping should always be preferred to resizing both during the training and test phase. Along this same direction, the no-down variant is very promising and suggests to work on full-resolution end-to-end processing to design better and more robust detectors, as also proposed in [37] for image forgery detection. More importantly, they shed some lights on the needs for well-designed evaluation protocols to assess the generalization capabilities of AI-generated image detectors in real-world scenarios.

9.10 Open Challenges

The advent of deep learning has given extraordinary impulse to both face manipulation methods and forensic detection tools. We have seen that successful detectors rely on inconsistencies at different levels, looking for both hidden and visible artifacts. One first important observation is that visual imperfections on faces will likely disappear soon. Newer GAN architectures [28] already improved upon this aspect by producing faces with even more details and highly realistic. Thus, relying exclusively on these traces could be a losing strategy in the long term. Turning to generic deep learning based-solutions, the main technical issue is probably the inability to adapt to situations not seen in the training phase. Misalignment between training and test, compression, and resizing are all sources of serious impairments and, at the same time, highly realistic scenarios for real-world applications. Also, to deal with the rapid advances in manipulation technology, deep networks should be able to adapt readily to new manipulations, without a full re-training, which may be simply impossible for lack of training data or entail catastrophic forgetting phenomena.

A more fundamental problem is the two-player nature of this research which is common to many security-related fields. In fact, detection algorithms must confront with the capacity of an adversary to fool them. This means that new solutions are needed in order to cope with unforeseen attacks. This applies to any type of classifier and is also very well known in forensics, where many counter-forensics methods have been proposed in the literature in order to better understand weaknesses of current approaches and help to improve them over time.

In the following, we analyze some works that have shown the vulnerabilities of GAN detectors to different types of threats.

- *Adding adversarial perturbations.* It is well known, from the object recognition field, that suitable slight perturbations can induce misclassification [50]. Following this path, in [7] it has been investigated the robustness of GAN detectors to imperceptible noise both in a white-box and in a black-box scenario. The authors show that it is possible to generate appropriate adversarial perturbations so as to misclassify fake images as real (see Fig. 9.14), but also the opposite. In addition, they show that the attack can survive JPEG compression. Interestingly, it is also possible to design an effective strategy in a black-box threat model when the adversary does not have perfect knowledge of the classifier but is aware about the type of classifier. A similar analysis is conducted in [19], where adversarial attacks are designed to fool co-occurrence-based GAN detectors.
- *Removing GAN fingerprints.* Instead of adding noise, one can take a different perspective and remove the specific fingerprints that are used to discriminate GAN images from real ones. This approach is pursued in [43], where an autoencoder-based strategy is proposed, that is trained using only real faces and is able to remove the high-frequency components that correspond to the fingerprints of the models used to generate synthetic images. At test time the autoencoder takes as input synthetic face images and modifies them so as to spoof GAN detection systems.

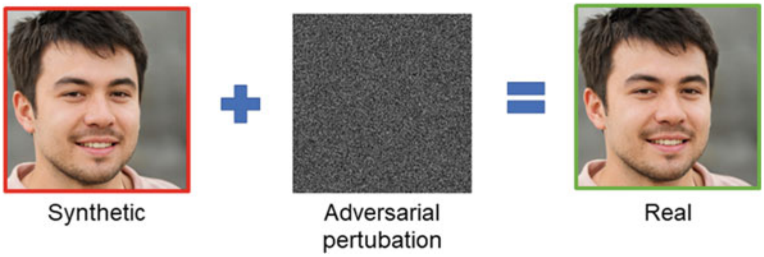


Fig. 9.14 A small and imperceptible adversarial perturbation can be added to the synthetic face image in order to fool the detector

- *Inserting camera fingerprints.* Another possible direction to attack GAN detectors is to insert the specific camera traces that characterize real images. In fact, real images are characterized by their own device and model fingerprints, as explained before. Such differences are important to carry out camera model identification from image content but can also be used to better highlight anomalies caused by image manipulations [52]. In [13] it is proposed a targeted black-box attack that is based on a GAN architecture, able to insert specific real camera traces in a synthetic images. In this way it is possible not only to fool a GAN detector without any prior information on its architecture, but also to fool a camera model identification algorithm, that will attribute the GAN image to the targeted camera under attack.

It is worth observing that all these approaches generate face images that are visually indistinguishable from real ones. This makes clear that a good GAN detector should always taken into account possible adversarial attacks and include proper strategies to face them. Another issue for forensics deep learning-based methods is interpretability. The black-box nature of these approaches makes it difficult to understand the reason behind a certain decision. Hence it is important to develop strategies that increase the level of understanding so as to improve its design and maybe also increase robustness to possible malicious attacks.

Overall, we can conclude that AI synthetic face detection is not a trivial task and, despite the huge effort made by the scientific community, we need to develop more reliable tools, that should also include anti-forensics and adversarial attacks since these techniques are widespread and can seriously impair the detection performance. It is difficult to forecast whether detection tools will be able to ensure a good defense against a bad use of synthetic content over the web or if active protection technology will become necessary. However, we believe that developing reliable detectors that possess good features in terms of generalization and robustness can represent a first step to protect our society.

Acknowledgements This material is based on research sponsored by the Defense Advanced Research Projects Agency (DARPA) and the Air Force Research Laboratory (AFRL) under agreement number FA8750-20-2-1004. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA and AFRL or the U.S. Government. This work is also supported by the PREMIER project, funded by the Italian Ministry of Education, University, and Research within the PRIN 2017 program and by a Google gift.

References

1. Albright M, McCloskey S (2019) Source generator attribution via inversion. In: IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW), pp 96–103
2. Barni M, Kallas K, Nowroozi E, Tondi B (2020) CNN detection of GAN-generated face images based on cross-band co-occurrences analysis. In: IEEE international workshop on information forensics and security (WIFS), pp 1–6
3. Berthelot D, Schumm T, Metz L (2017) BEGAN: boundary equilibrium generative adversarial networks. arXiv preprint arXiv:1703.10717
4. Bonettini N, Bestagini P, Milani S, Tubaro S (2020) On the use of Benford's law to detect GAN-generated images. In: IEEE international conference on pattern recognition
5. Boroumand M, Chen M, Fridrich J (2019) Deep residual network for steganalysis of digital images. IEEE Trans Inform Forensics Secur 14(5):1181–1193
6. Brock A, Donahue J, Simonyan K (2018) Large scale GAN training for high fidelity natural image synthesis. In: International conference on learning representations (ICLR)
7. Carlini N, Farid H (2020) Evading deepfake-image detectors with white- and black-box attacks. In: IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW), pp 2804–2813
8. Chai L, Bau D, Lim SN, Isola P (2020) What makes fake images detectable? Understanding properties that generalize. In: European conference on computer vision (ECCV). Springer, pp 103–120
9. Chen M, Fridrich J, Goljan M, Lukás J (2008) Determining image origin and integrity using sensor noise. IEEE Trans Inform Forensics Secur 3(1):74–90
10. Choi Y, Choi M, Kim M, Ha JW, Kim S, Choo J (2018) StarGAN: unified generative adversarial networks for multi-domain image-to-image translation. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 8789–8797
11. Chollet F (2017) Xception: Deep learning with depthwise separable convolutions. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR)
12. Cozzolino D, Poggi G, Verdoliva L (2017) Recasting residual-based local descriptors as convolutional neural networks: an application to image forgery detection. In: ACM workshop on information hiding and multimedia security, pp 1–6 (2017)
13. Cozzolino D, Thies J, Roessler A, Niessner M, Verdoliva L (2019) SpoC: spoofing camera fingerprints. In: IEEE CVPR Workshops, June 2021
14. Cozzolino D, Thies J, Rössler A, Riess C, Nießner M, Verdoliva L (2018) ForensicTransfer: weakly-supervised domain adaptation for forgery detection. arXiv preprint arXiv:1812.02510
15. Dang H, Liu F, Stehouwer J, Liu X, Jain A (2020) On the detection of digital face manipulation. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 5781–5790
16. Durall R, Keuper M, Keuper J (2020) Watch your up-convolution: CNN based generative deep neural networks are failing to reproduce spectral distributions. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR)

17. Dzanic T, Shah K, Witherden F (2020) Fourier spectrum discrepancies in deep network generated images. In: Conference on neural information processing systems (NeurIPS)
18. Frank J, Eisenhofer T, Schönherr L, Fischer A, Kolossa D, Holz T (2020) Leveraging frequency analysis for deep fake image recognition. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR)
19. Goebel M, Manjunath B (2020) Adversarial attacks on co-occurrence features for GAN detection. arXiv preprint arXiv:2009.07456
20. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. In: Conference in neural information processing systems (NIPS)
21. Gragnaniello D, Cozzolino D, Marra F, Poggi G (2021) Verdoliva L (2021) Are GAN generated images easy to detect? IEEE international conference on multimedia and expo (ICME), A critical analysis of the state-of-the-art. In
22. Han X, Ji Z, Wang W (2020) Low resolution facial manipulation detection. In: IEEE international conference on visual communications and image processing (VCIP), pp 431–434
23. Hu S, Li Y, Lyu S (2020) Exposing GAN-generated faces using inconsistent corneal specular highlights. In: IEEE international conference on acoustics, speech and signal processing 2021
24. Jain A, Majumdar P, Singh R, Vatsa M (2020) Detecting GANs and retouching based digital alterations via DAD-HCNN. In: IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW), pp 2870–2879
25. Karras T, Aila T, Laine S, Lehtinen J (2018) Progressive growing of GANs for improved quality, stability, and variation. In: International conference on learning representations (ICLR)
26. Karras T, Aittala M, Hellsten J, Laine S, Lehtinen J, Aila T (2020) Training generative adversarial networks with limited data. In: Conference on neural information processing systems (NeurIPS)
27. Karras T, Laine S, Aila T (2019) A style-based generator architecture for generative adversarial networks. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 4396–4405
28. Karras T, Laine S, Aittala M, Hellsten J, Lehtinen J, Aila T (2020) Analyzing and improving the image quality of StyleGAN. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 8110–8119
29. Khodabakhsh A, Ramachandra R, Raja K, Wasnik P, Busch C (2018) Fake face detection methods: Can they be generalized? In: International conference of the biometrics special interest group (BIOSIG), pp 1–6
30. Kingma DP, Dhariwal P (2018) Glow: generative flow with invertible 1 × 1 convolutions. In: Conference on neural information processing systems (NIPS), pp 10236–10245
31. Li H, Li B, Tan S, Huang J (2020) Detection of deep network generated images using disparities in color components. Signal Process 174
32. Liu M, Ding Y, Xia M, Liu X, Ding E, Zuo W, Wen S (2019) STGAN: a unified selective transfer network for arbitrary image attribute editing. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 3673–3682
33. Liu Z, Luo P, Wang X, Tang X (2015) Deep learning face attributes in the wild. In: International conference on computer vision (ICCV)
34. Lukas J, Fridrich J, Goljan M (2006) Digital camera identification from sensor pattern noise. IEEE Trans Inform Forensics Secur 1(2):205–214
35. Marra F, Gragnaniello D, Cozzolino D, Verdoliva L (2018) Detection of GAN-generated fake images over social networks. In: IEEE conference on multimedia information processing and retrieval (MIPR)
36. Marra F, Gragnaniello D, Verdoliva L, Poggi G (2019) Do GANs leave artificial fingerprints? In: IEEE conference on multimedia information processing and retrieval (MIPR), pp 506–511
37. Marra F, Gragnaniello D, Verdoliva L, Poggi G (2020) A full-image full-resolution end-to-end-trainable CNN framework for image forgery detection. IEEE Access 8
38. Marra F, Saltori C, Boato G, Verdoliva L (2019) Incremental learning for the detection and classification of GAN-generated images. In: IEEE international workshop on information forensics and security (WIFS)

39. Matern F, Riess C, Stamminger M (2019) Exploiting visual artifacts to expose deepfakes and face manipulations. In: IEEE winter applications of computer vision workshops (WACVW), pp 83–92
40. McCloskey S, Albright M (2019) Detecting GAN-generated imagery using saturation cues. In: IEEE international conference on image processing (ICIP)
41. Mi Z, Jiang X, Sun T, Xu K (2020) GAN-generated image detection with self-attention mechanism against GAN generator defect. IEEE J Sel Top Signal Process 14(5):969–981
42. Nataraj L, Mohammed T, Manjunath B, Chandrasekaran S, Flenner A, Bappy J, Roy-Chowdhury A (2019) Detecting GAN generated fake images using co-occurrence matrices. In: IS&T Electronic imaging, media watermarking, security, and forensics, pp 532–1–532–7
43. Neves JC, Tolosana R, Vera-Rodriguez R, Lopes V, Proena H, Fierrez J (2020) GANprintR: improved fakes and evaluation of the state of the art in face manipulation detection. IEEE J Sel Top Signal Process 14(5):1038–1048
44. Park T, Liu MY, Wang T-C, Zhu J-Y (2019) Semantic image synthesis with spatially-adaptive normalization. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 2337–2346
45. Pei K, Cao Y, Yang J, Jana S (2017) Deepxplore: automated whitebox testing of deep learning systems. In: 26th symposium on operating systems principles, pp 1–18
46. Perarnau G, van de Weijer J, Raducanu B, Álvarez J (2016) Invertible conditional GANs for image editing. In: NIPS workshop on adversarial training
47. Qian Y, Yin G, Sheng L, Chen Z, Shao J (2020) Thinking in frequency: face forgery detection by mining frequency-aware clues. In: European conference on computer vision (ECCV), pp 86–103
48. Rössler A, Cozzolino D, Verdoliva L, Riess C, Thies J, Nießner M (2019) FaceForensics++: learning to detect manipulated facial images. In: International conference on computer vision (ICCV)
49. Shen Y, Luo P, Yan J, Wang X, Tang X (2018) Faceid-GAN: learning a symmetry three-player GAN for identity-preserving face synthesis. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 821–830
50. Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R (2014) Intriguing properties of neural networks. In: International conference on learning representations (ICLR)
51. Upchurch P, Gardner J, Pleiss G, Pless R, Snavely N, Bala K, Weinberger K (2017) Deep feature interpolation for image content changes. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 7064–7073
52. Verdoliva L (2020) Media forensics and deepfakes: an overview. IEEE J Sel Top Signal Process 14(5):910–932
53. Wang R, Juefei-Xu F, Ma L, Xie X, Huang Y, Wang J, Liu Y (2020) FakeSpotter: a simple yet robust baseline for spotting AI-synthesized fake faces. In: International joint conference on artificial intelligence (IJCAI), pp 3444–3451
54. Wang SY, Wang O, Zhang R, Owens A, Efros A (2020) CNN-generated images are surprisingly easy to spot... for now. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR)
55. Wu PW, Lin YJ, Chang H-C, Chang E, Liao SW (2019) RelGAN: multi-domain image-to-image translation via relative attributes. In: IEEE international conference on computer vision (ICCV)
56. Xuan X, Peng B, Wang W, Dong J (2019) On the generalization of GAN image forensics. In: Chinese conference on biometric recognition, pp 134–141
57. Yang X, Li Y, Qi H, Lyu S (2019) Exposing GAN-synthesized faces using landmark locations. In: ACM workshop on information hiding and multimedia security, pp 113–118
58. Yao G, Yuan Y, Shao T, Zhou K (2020) Mesh guided one-shot face reenactment using graph convolutional networks. In: ACM international conference on multimedia, pp 1773–1781
59. Yousfi Y, Butora J, Khvedchenya E, Fridrich J (2020) ImageNet pre-trained CNNs for JPEG Steganalysis. In: IEEE international workshop on information forensics and security (WIFS), pp 1–6

60. Yu N, Davis L, Fritz M (2019) Attributing fake images to GANs: learning and analyzing GAN fingerprints. In: IEEE international conference on computer vision (ICCV)
61. Zhang X, Karaman S, Chang SF (2019) Detecting and simulating artifacts in GAN fake images. In: IEEE international workshop on information forensics and security (WIFS), pp 1–6
62. Zhu JY, Park T, Isola P, Efros A (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. In: IEEE international conference on computer vision (ICCV), pp 2223–2232