

Identification of GAN images fingerprints

Signal, Image and Video [145858]

University of Trento
Master of Artificial Intelligence Systems

Ali AKAY
Mert AKKOR

Instructors: Francesco De Natale, Giulia Boata and Cecilia Pasquini

1 Introduction

Generative adversarial networks(GAN) has demonstrated immense promise for a range of applications and associated areas of computer vision in the past few years. With the current pace of progress, it is a sure thing that they will soon be able to create high-quality, practically indistinguishable photographs and videos from actual ones.

While recent state-of-the-art visual forensics techniques demonstrate impressive results for detecting fake visual media, they have only focused on semantic, physical, or statistical inconsistency of specific forgery scenarios, e.g., copy-move manipulations or face swapping [1]. Forensics on GAN-generated images shows good accuracy, but each method operates on only one GAN architecture by identifying its unique artifacts and results deteriorate when the GAN architecture is changed. It is still an open question of whether GANs leave stable marks that are commonly hared by their generated images. That motivates us to investigate an effective feature representation that differentiates GAN-generated images from real ones.

1.1 Problem of GAN Detection

Unfortunately, realistic images created by GAN pose serious security risks. The future influence of this technology in the wrong hands is of serious concerns. The already worrying phenomena of false news adds more strength to well-crafted fake multimedia, and there is an immediate need for multimedia forensic countermeasures [2]. While today's GAN-based manipulations frequently present objects that lift the observer suspect, see Fig.1(top), this is not always the case (bottom), and it is only a matter of

time before visual inspection is regularly passed on by GAN-generated images. To detect such fakes, suitable multimedia forensic tools are therefore needed.



Figure 1: Sample images generated by Pro-GAN (a), Cycle-GAN (b), Star-GAN. Top: easily detected bad results. Bottom: photorealistic results.

A significant number of approaches have been proposed in recent years to detect fake visual data, based on their semantic, physical, or statistical anomalies. GAN challenges to visual forensics. There is a widespread concern about the impact of this technology when used maliciously. This issue has also received increasing public attention, in terms of disruptive consequences to visual security, laws, politics, and society in general [1]. Therefore, it is critical to look into effective visual forensics against threats from GANs.

1.2 Goal of the project

In this work, we show that in the images it generates, each GAN leaves its particular fingerprint, just as real-world cameras identify acquired images with traces of their pattern. In fact, each individual device, due to manufacturing imperfections, leaves a unique and stable mark on each acquired photo, the photo-response non-uniformity (PRNU) pattern. Experiments on source recognition with many common GANs show such fingerprints to be a valuable asset for forensic analysis. Our studies with several common GAN architectures (StyleGAN and StyleGAN2) and datasets (FFHQ) demonstrate that GAN leaves unique fingerprints on the image they create that can be used to conduct accurate forensic analysis.

2 PRNU-Based Analysis

PRNU-based forgery detection was first proposed in [4] based on two steps: i) the camera PRNU pattern is estimated off-line from a large number of images taken from the camera, and ii) the target image PRNU is estimated at test time, by means of a denoising filter, and compared with the reference. In this project we use the technique for the GAN images.(see Figure 2).

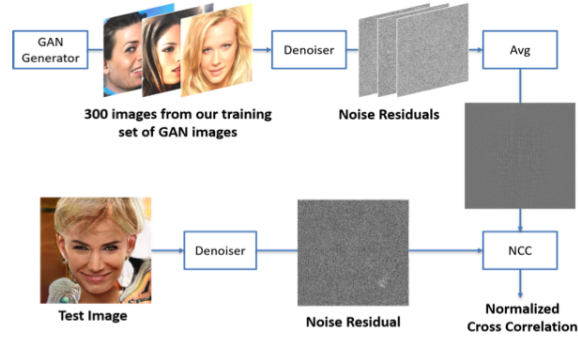


Figure 2: PRNU-Based GAN Fingerprint

Images noise pattern can be estimated from the noise residual w of an image x . Our approach to PRNU-based GAN image detection measures the similarity between a reference fingerprints \mathbf{c} and the noise residuals \mathbf{w} of an image by using **Wiener Filter** we obtain an estimate of the PRNU in order to solve the identification of the image under investigation in terms of the normalized cross correlation [2].

$$\text{norm_corr}(\mathbf{k}, \mathbf{w}) = \frac{\sum_{n=0}^{n-1} k[n] * w[n]}{\sqrt{\sum_{n=0}^{n-1} k[n]^2 * \sum_{n=0}^{n-1} w[n]^2}}$$

Most of our experiments confirm that normal images fingerprints has the higher correlation between each other over a 500 normal pictures residuals.(See Figure 3)

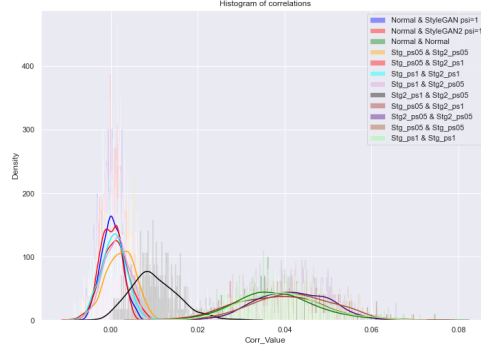


Figure 3: Histogram of Correlation between Normal-Generated Images fingerprints and residuals

Figure 3 shows the histogram of same type of images (eg. green) and between each other correlations. Normal-GAN correlations are distributed around zero, indicating no correlations between generated images and unrelated normal images fingerprints. On the contrary, same-GAN and Normal-Normal correlations are markedly larger than zero. Moreover, the distributions are well separated. These findings provide a reasonable answer to our fundamental question, demonstrating that each GAN leaves a distinctive mark that can reasonably be called a fingerprint on each image generated by it.

3 Experiments

In this section, we assess the performance of the our Image identification method. We labeled 0 and 1 from 500 Generated Images and 500 Normal Images accordingly. We put correlation threshold to find the best separability between GAN and Normal pictures (see Figure 4). Then we obtained 500 generated image fingerprints and 500 normal image fingerprints and compared each of them with normal image residuals. (see Table 1).

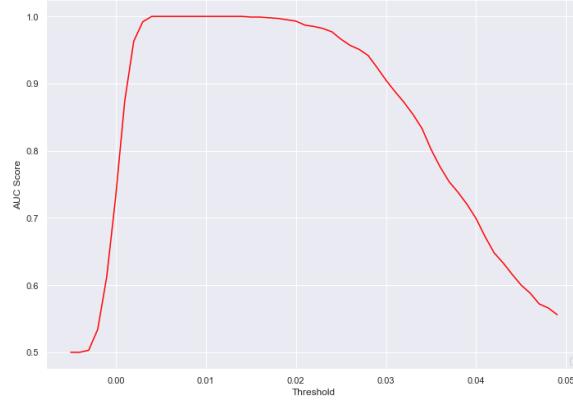


Figure 4: AUC Scores by Correlation Thresholds

We performed a threshold tuning, between 0 and 0.01 for all 1000(500 generated and 500 normal image) test images. Results of the area under the ROC curve from predictions can be seen from the graph(Figure 4). According to the AUC graph (Figure 4) we should have select the threshold as 0.005.

3.1 Dataset for testing

For each image type we take 300 test images. Then, we extract fingerprints and residuals for each image type. Following this, we applied our classification model according to threshold value.

Image Type	Number of Images
StyleGAN psi = 0.5	300
StyleGAN psi = 1	300
StyleGAN2 psi = 0.5	300
StyleGAN2 psi = 1	300
FFHQ Dataset (Normals)	300

Table 1: Dataset overview

3.2 Results

Here it can be seen that, our classification works significantly with 0.005 correlation threshold for our dataset (see Table 1). Also in Figure 3, when we look at the intersection point between generated images and normal images it shows the right threshold, so that it proves our classification.

Threshold Value	StyleGAN psi05	StyleGAN psi1	StyleGAN2 psi05	StyleGAN2 psi1	Normal
0.001	0.59	0.65	0.54	0.59	0.58
0.005	0.96	0.97	0.95	0.98	0.97

Table 2: Accuracy Results






Image	Correlation with Normal Fingerprints	Image Type	Prediction
	-0.000859	StyleGAN psi05	1
	-0.001782	StyleGAN psi1	1
	0.000556	StyleGAN2 psi05	1
	-0.006004	StyleGAN2 psi1	1
	0.001506	Normal Images	0

Table 3: Table 3 shows the classification examples from all generated image types.

4 Conclusion

Artificial intelligence has largely changed the rules of virtual security. High- quality fakes now seem to come out from an assembly line calling for an extraordinary effort on part of both scientists and policymakers. In fact, today’s multimedia forensics is in full development, major agencies are funding large research initiatives, and scientists from many different field are contributing actively, with fast advances in ideas and tools [4].

In this report, we have analysed PRNU based GAN image identification. Our experiments show that even a small difference in GAN training (e.g., the difference in initialization) can leave a distinct fingerprint that commonly exists over all its generated images. These fingerprints can be essential to detect malicious activities and identify the criminals in order to keep multimedia environments safe, additionally very important improvement for cyber-security area.

5 Future works

As for future work, most significant errors on GAN pictures are located on people's hair, mouth, eyes, teeth and image background. According to further studies and European Conference on Computer Vision, patch-classifiers are more efficient and accurate. It takes a part of an image and operates locally, not on whole image unlike our method. Further more, PRNU based identification can be applied to GAN generated videos.

References

- [1] Ning Yu, Larry Davis and Mario Fritz *Attributing Fake Images to GANs: Learning and Analyzing GAN Fingerprints*.
- [2] Francesco Marra, Diego Gragnaniello, Luisa Verdoliva and Giovanni Poggi *Do GANs leave artificial fingerprints?*.
- [3] Lucy Chai, David Bau, Ser-Nam Lim and Phillip Isola *What makes fake images detectable? Understanding properties that generalize*.
- [4] Luisa Verdoliva *Media Forensics and DeepFakes:an overview*.
- [5] Jessica Fridrich *Digital Image Forensics Using Sensor Noise*.