

Do GANs leave artificial fingerprints?

Francesco Marra, Diego Gragnaniello, Luisa Verdoliva, Giovanni Poggi
DIETI – University Federico II of Naples
Via Claudio 21, 80125 Napoli – ITALY
{francesco.marra, diego.gragnaniello, verdoliv, poggi}@unina.it

Abstract

In the last few years, generative adversarial networks (GAN) have shown tremendous potential for a number of applications in computer vision and related fields. With the current pace of progress, it is a sure bet they will soon be able to generate high-quality images and videos, virtually indistinguishable from real ones. Unfortunately, realistic GAN-generated images pose serious threats to security, to begin with a possible flood of fake multimedia, and multimedia forensic countermeasures are in urgent need. In this work, we show that each GAN leaves its specific fingerprint in the images it generates, just like real-world cameras mark acquired images with traces of their photo-response non-uniformity pattern. Source identification experiments with several popular GANs show such fingerprints to represent a precious asset for forensic analyses.

1 Introduction

Generative adversarial networks are pushing the limits of image manipulation. A skilled individual can easily generate realistic images sampled from a desired distribution [19, 8, 1], or convert original images to fit a new context of interest [21, 9, 23, 12, 3]. With progressive GANs [10], images of arbitrary resolution can be created, further improving the level of photorealism.

There is widespread concern on the possible impact of this technology in the wrong hands. Well-crafted fake multimedia add further momentum to the already alarming phenomenon of fake news, if “seeing is believing”, as they say. Although today’s GAN-based manipulations present often artifacts that raise the suspect of observers, see Fig.1(top), this is not always the case (bottom), and it is only a matter of time before GAN-generated images will consistently pass visual scrutiny. Therefore, suitable multimedia forensic tools are required to detect such fakes.

In recent years, a large number of methods have been proposed to single out fake visual data, relying on their se-

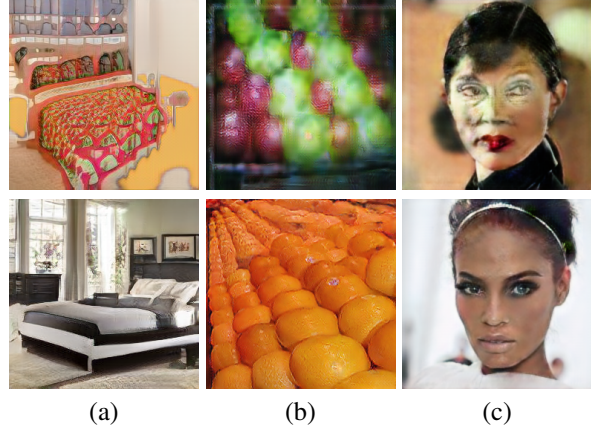


Figure 1. Sample images generated by Pro-GAN (a), Cycle-GAN (b), Star-GAN. Top: easily detected bad results. Bottom: photorealistic results.

mantic, physical, or statistical inconsistencies [7].

Statistical-based approaches, in particular, rely on the long trail of subtle traces left in each image by the acquisition devices, traces that can be hardly disguised even by a skilled attacker. 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 [13], which can be estimated and used as a sort of *device* fingerprint. Likewise, each individual acquisition model, due to its peculiar in-camera processing suite (demosaicking, compression, etc.), leaves further model-related marks on the images, which can be used to extract a *model* fingerprint [5]. Such fingerprints can be used to perform image attribution [13, 2], as well as to detect and localize image manipulations [2, 5], and represent one of the strongest tools in the hands of the forensic analyst.

GANs have little in common with conventional acquisition devices, and GAN-generated images will not show the same camera-related marks. Still, they are the outcome of complex processing systems involving a large number of filtering processes, which may well leave their own distinc-

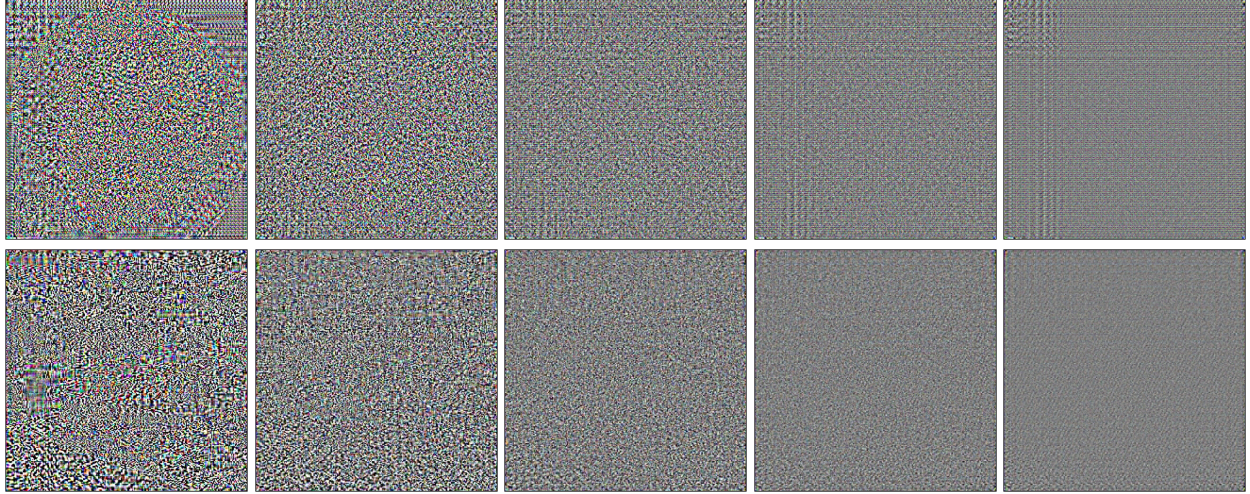


Figure 2. Cycle-GAN o2a (top) and Pro-GAN kitchen (bottom) fingerprints estimated with 2, 8, 32, 128, 512 residuals.

tive marks on output images. So the question¹ is: do GANs leave artificial fingerprints? That is, do the images generated by a given GAN share a common and stable pattern that allows to establish their origin? And, if this is the case, how reliable will such a fingerprint be? How robust to defensive measures? And how discriminative about the image origin?

In this paper we investigate on this interesting issue, and provide a first answer to the above questions. Our experiments with several popular GAN architectures and datasets, show that GAN do leave specific fingerprints on the image they generate, which can be used to carry out reliable forensic analyses.

2 Related Work

Recently there has been a growing interest in distinguishing GAN-generated images from real ones. As shown in Fig.1, the current state of the art in GANs is far from perfection, and often generated images exhibit strong visual artifacts that can be exploited for forensic use. For example, to detect fake faces, [15] exploits visual features regarding eyes, teeth and facial contours. Tellingly, the authors observe that in GAN-generated images the color of left and right eye are often inconsistent. Color information is also used in [16, 11]. In particular, [16] proposes to use some features shared by different GAN architectures, related to the synthesis of RGB color channels. Other methods rely on deep learning. Several architectures have been tested so far [14, 17, 20] showing a good accuracy in detecting GAN-generated images, even on compressed images. Unfortunately, if a network is trained on a specific architecture,

its performance degrades sharply when used to detect image generated by another architecture [4]. This observation suggests the presence of different artifacts peculiar of each specific GAN model. Recently, it has also been shown [22] that a deep network can reliably discriminate images generated with different architectures. However, the network requires intensive training on an aligned dataset, and there is no hint, let alone exploitation, of the presence of GAN-induced fingerprints.

3 Exposing GAN fingerprints

In this Section we show evidence on the existence of GAN fingerprints. This goal is pursued in a minimal experimental setting, considering only two GANs, a Cycle-GAN trained to convert orange images into apple images and a Progressive-GAN (Pro-GAN) trained to generate kitchen images, call them GAN A and B, from now on. Lacking any statistical model, we consider an extraction pipeline similar to that of the PRNU pattern. For the generic image X_i generated by a given GAN, the fingerprint represents a disturbance, unrelated with the image semantics. Therefore, we first estimate the high-level image content, $\hat{X}_i = f(X_i)$, through a suitable denoising filter $f(\cdot)$, then subtract it from the original image to extract the noise residual

$$R_i = X_i - f(X_i) \quad (1)$$

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

$$R_i = F + W_i \quad (2)$$

¹Winking at P.K.Dick novel: “Do androids dream of electric sheep?”

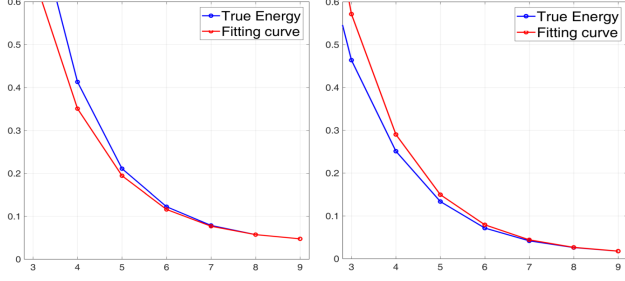


Figure 3. True energy and fitting curve for the Cycle-GAN and Pro-GAN fingerprints of Figure 2.

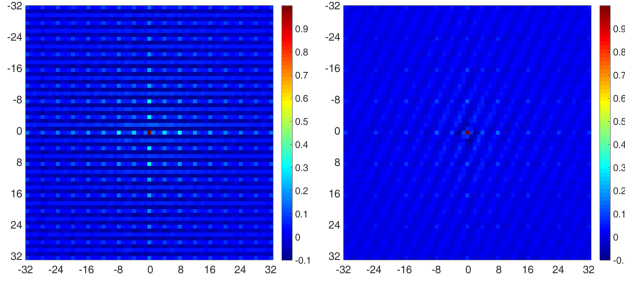


Figure 4. Autocorrelation matrices of the Cycle-GAN and Pro-GAN fingerprints ($N=512$) of Figure 2.

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

$$\hat{F} = \frac{1}{N} \sum_{i=1}^N R_i \quad (3)$$

Fig.2 shows (suitably amplified) the fingerprints of the two GANs, estimated over a growing number of residuals, $N = 2, 8, 32, 128, 512$. Of course, for low values on N , the estimates are dominated by image-related noise. However, as N grows, the additive noise component tends to vanish and both estimates converge to stable quasi-periodical patterns, which we regard as accurate approximations of the true GAN fingerprints. In Fig.3 we show the energy $E(N)$ of these estimated fingerprints as a function of N , together with the best fitting curve of the type

$$\hat{E}(N) = E_\infty + E_0 \times 2^{-N} \quad (4)$$

The fitting is clearly very accurate for large values of N , and the E_∞ value estimates the energy of the limit fingerprint, 0.0377 and 0.0088, respectively. Fig.4, instead, shows the autocorrelation functions of the two estimates for $N=512$, with clear quasi-periodical patterns providing further evidence of the non-random nature of these signals.

We now take a more application-oriented point of view, looking for these fingerprints' ability to tell apart images of

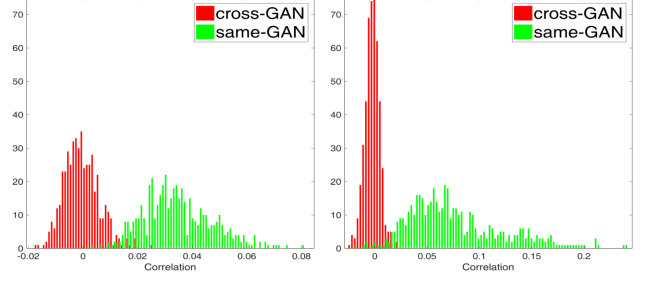


Figure 5. Correlation of Cycle-GAN (left) and Pro-GAN (right) residuals with same/cross-GAN fingerprints.

different origin. Based on image-to-fingerprint correlation or similar indicators, meaningful fingerprints should allow one to decide which of the two GANs generated a given image.

Let

$$\text{corr}(X, Y) = \tilde{X} \odot \tilde{Y} \quad (5)$$

be the correlation index between images X and Y , where \tilde{X} is the zero-mean unit-norm version of X and \odot indicates inner product. For both GANs under analysis, we regard the estimates obtained with $N = 2^9$ as the “true” fingerprints, F_A and F_B , respectively. Then, we compute the correlation indices between residuals $R_i^A, i = 1, \dots, M$ generated by GAN A (and not used to estimate the fingerprint), and the same-GAN (F_A) and cross-GAN (F_B) fingerprints, that is

$$\rho_{i,\text{same}}^A = \text{corr}(R_i^A, F_A) \quad (6)$$

and

$$\rho_{i,\text{cross}}^A = \text{corr}(R_i^A, F_B) \quad (7)$$

Fig.5(left) shows the histograms of same-GAN (green) and cross-GAN (red) correlations. Cross-GAN correlations are evenly distributed around zero, indicating no correlation between generated images and the unrelated fingerprint. On the contrary, same-GAN correlations are markedly larger than zero, 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, see Fig.5(right). Moreover, in both cases the two distributions are well separated, allowing reliable discrimination. The corresponding receiver operating curves (ROC) are nearly perfect with area under curve (AUC) 0.990 and 0.998, respectively.

We carried out similar experiments for many other GANs, differing for architecture and/or training set, obtaining always similar results. These results provide a convincing answer to our fundamental question, showing that each GAN leaves a distinctive mark on each image generated by it, which can be legitimately called fingerprint.

4 Source identification experiments

Let us now consider a more challenging experimental setting, to carry out larger-scale source identification tests. We consider three GAN architectures, Cycle-GAN, Pro-GAN, and Star-GAN. Cycle-GAN was proposed in [23] to perform image-to-image translation. The generator takes an input image of the source domain and transforms it into a new image of the target domain (*e.g.*, apples to oranges). To improve the photorealism of generated images, a cycle consistency constraint is enforced. Here, we consider several Cycle-GAN networks, trained by the authors on different source/target domains. The second architecture, Progressive-GAN [10], uses progressively growing generator and discriminator to create images of arbitrary-size which mimic images of the target domain. In this case too, six different target domains are considered. Like Cycle-GAN, Star-GAN [3] performs image-to-image translation, but adopts a unified approach such that a single generator is trained to map an input image to one of multiple target domains, which can be selected by the user. By sharing the generator weights among different domains, a dramatic reduction of the number of parameters is achieved. Finally, we include also two sets, from the RAISE dataset [6], of images acquired by real cameras, so as to compare the behavior of real-world and GAN fingerprints. Table I lists all networks and cameras, with corresponding abbreviations. For each dataset \mathcal{A} , we generate/take 1000 RGB images of 256×256 pixels, extract residuals, and use $N=512$ of them to estimate the fingerprint F_A , and the remaining $M=488$, $\{R_1^A, \dots, R_M^A\}$ for testing.

Architecture	Target / Camera model	Abbreviation
Cycle-GAN	apple2orange	C1
	horse2zebra	C2
	monet2photo	C3
	orange2apple	C4
	photo2Cezanne	C5
	photo2Monet	C6
	photo2Ukiyoe	C7
	photo2VanGogh	C8
	zebra2horse	C9
Pro-GAN	bedroom	G1
	bridge	G2
	church	G3
	kitchen	G4
	tower	G5
	celebA	G6
Star-GAN	black hair	S1
	blond hair	S2
	brown hair	S3
	male	S4
	smiling	S5
n/a	Nikon-D90	N1
	Nikon-D7000	N2

Table 1. Cameras and GANs used in the experiments

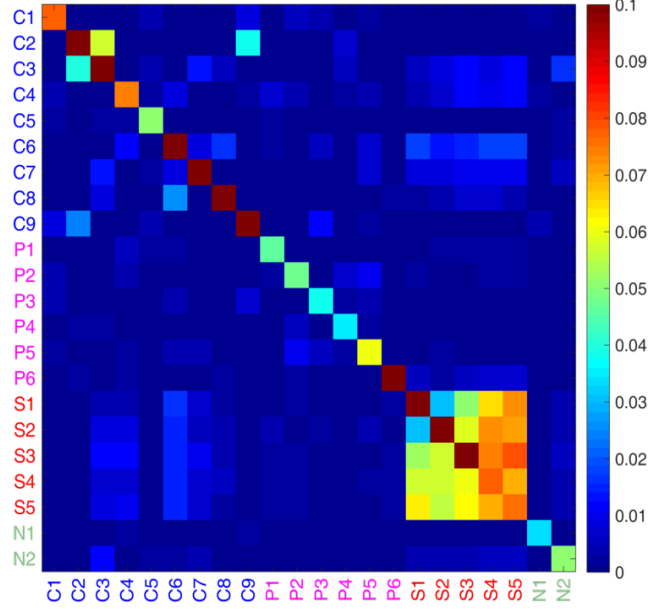


Figure 6. Average residual-fingerprint correlations.

First of all we compute the average correlation between all sets of residuals and all fingerprints, that is

$$\langle \rho \rangle_B^A = \frac{1}{M} \sum_{i=1}^M \text{corr}(R_i^A, F_B) \quad (8)$$

with A, B spanning all sets. Fig.6 shows a false-color representation of all such correlations. It appears that diagonal entries are much larger, in general, than off-diagonal ones, confirming that residuals of a dataset correlate well only with the fingerprint of the same dataset, be it GAN or natural. There is also a clear block structure, showing that some (weaker) correlation exists between residuals of a dataset and fingerprints of “sibling” datasets, as associated with the same GAN architecture. This holds especially for the Star-GAN datasets, since the weights of a single generator are shared among all target domains. Finally, as expected, no significant correlation exists between real and GAN-generated images, which can be told apart easily based on their respective fingerprints.

We now perform camera attribution. For each image, we compute the distance between the corresponding residual and all fingerprints, attributing the image with a minimum-distance rule. In Fig.7 we show the resulting ROCs, and in Fig.8 the confusion matrix (entries below 1% are canceled to improve readability). Despite the $2 \times$ zooming, the ROC figure is hard to read as all curves amass in the upper-left corner. On the other hand, this is the only real message we wanted to gather from this figure: attribution is very accurate in all cases, with the only exception of the Star-GAN male and smiling networks. This observation is reinforced

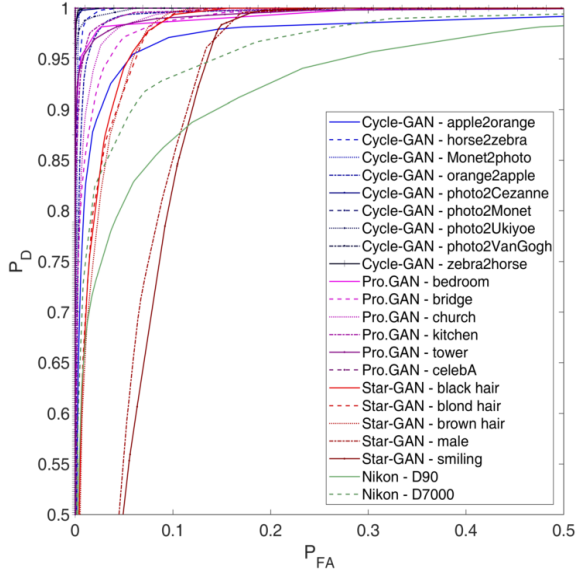


Figure 7. One-vs.-all source identification ROCs.

by the confusion matrix, showing almost perfect attribution in all cases (with the same exceptions as before), and with a slightly worse performance for the real cameras, characterized by a lower-energy PRNU. Since real cameras usually compress images at high quality to save storage space, we also repeated the attribution experiment after JPEG compressing all GAN-generated images at QF=95, observing a negligible loss in accuracy, from 90.3% to 90.1%.

We conclude this Section by reporting very briefly on the results obtained in the “Forensics GAN Challenge” [18] organized in June-July 2018 by the US National Institute of Standards and Technology in the context of the Medifor program. The goal was to classify as real or GAN-generated 1000 images of widely different resolution, from 52×256 to 4608×3072 pixels. As baseline method we used a deep network trained on a large number of images retrieved from the InterNet. However, we also tested the GAN fingerprint idea, following the scheme outlined in Fig.9. We computed fingerprints for several popular GANs and, eventually, identified a large cluster of size- 1024×1024 images generated with the same GAN. This allowed us to improve the deep net accuracy by a simple fusion rule, for a final 0.999 AUC.

5 Conclusions and future work

The goal of this work was to demonstrate the existence of GAN fingerprints and their value for reliable forensic analyses. We believe both facts are supported by a sufficient experimental evidence. This answers our fundamental question, but introduces many more questions and interesting topics which deserve further investigation.

First of all, it is important to understand how the fingerprint depends on the network, both its architecture (number and type of layers) and its specific parameters (filter weights). This may allow one to improve the fingerprint quality or, with the attacker’s point of view, find ways to remove the fingerprint from generated images as a counter-forensic measure. Along the same path, our preliminary results suggest that training the same architecture with different datasets gives rise to well distinct fingerprints. Is this true in general? Will fine-tuning produce similar effects?

Under a more practical point of view, further studies are necessary to assess the potential of GAN fingerprints in multimedia forensics. Major aims, besides source identification, are the discrimination between real and GAN-generated images, and the localization of GAN-generated material spliced in real images. It is also important to study the robustness of such fingerprints to subsequent processing, such as JPEG compression, resizing, blurring, noising. Finally, it is worth assessing the dependence of performance on the number and size of images used for fingerprint estimation, with blind attribution and clustering as an interesting limiting case.

References

- [1] D. Berthelot, T. Schumm, and L. Metz. BEGAN: Boundary Equilibrium Generative Adversarial Networks. *arXiv preprint: abs/1703.10717*, 2017.
- [2] M. Chen, J. Fridrich, M. Goljan, and J. Lukáš. Determining image origin and integrity using sensor noise. *IEEE Transactions on Information Forensics and Security*, 3:74–90, 2008.
- [3] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo. StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
- [4] D. Cozzolino, J. Thies, A. Rössler, C. Riess, and L. Verdoliva. ForensicTransfer: Weakly-supervised Domain Adaptation for Forgery Detection. *arXiv preprint arXiv:1812.02510*, 2018.
- [5] D. Cozzolino and L. Verdoliva. Noiseprint: a CNN-based camera model fingerprint. *arXiv preprint arXiv:1808.08396*, 2018.
- [6] D. Dang-Nguyen, C. Pasquini, V. Conotter, and G. Boato. RAISE: a raw images dataset for digital image forensics. In *Proc. of the 6th ACM Multimedia Systems Conference*, pages 219–224, 2015.
- [7] H. Farid. *Photo Forensics*. The MIT Press, 2016.
- [8] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville. Improved Training of Wasserstein GANs. In *Advances in Neural Information Processing Systems*, pages 5767–5777, 2017.
- [9] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

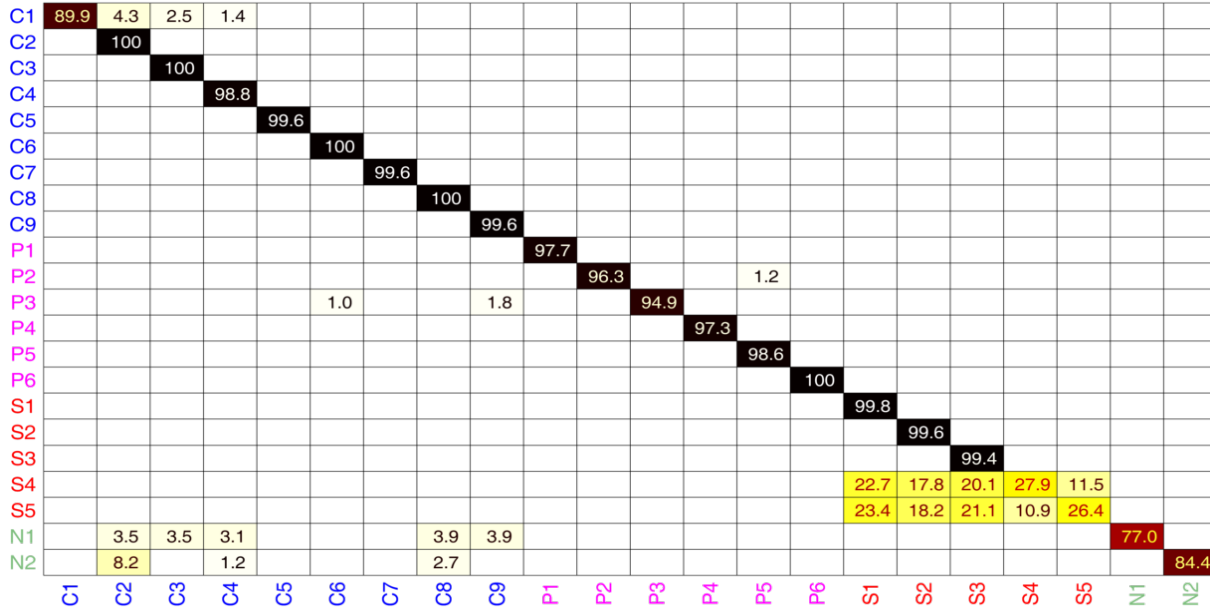


Figure 8. Source identification confusion matrix. Entries below 1% are canceled.

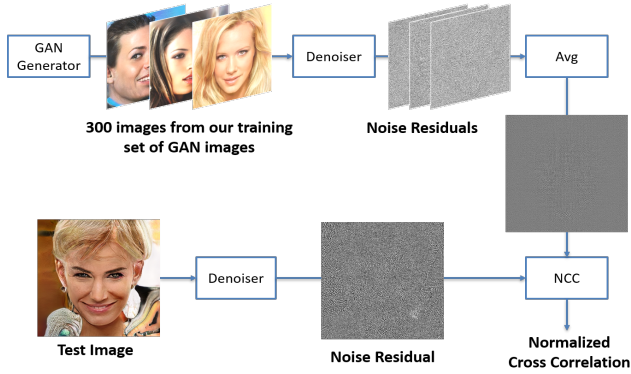


Figure 9. GAN-fingerprints for the Forensics Challenge.

[10] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *International Conference on Learning Representations*, 2018.

[11] H. Li, B. Li, S. Tan, and J. Huang. Detection of deep network generated images using disparities in color components. *arXiv preprint arXiv:1808.07276v1*, 2018.

[12] M.-Y. Liu, T. Breuel, and J. Kautz. Unsupervised image-to-image translation networks. In *Neural Information Processing Systems*, 2017.

[13] J. Lukáš, J. Fridrich, and M. Goljan. Digital camera identification from sensor pattern noise. *IEEE Transactions on Information Forensics and Security*, 1(2):205–214, 2006.

[14] F. Marra, D. Gragnaniello, G. Poggi, and L. Verdoliva. Detection of GAN-Generated Fake Images over Social Networks. In *IEEE Conference on Multimedia Information Processing and Retrieval*, pages 384–389, April 2018.

[15] F. Matern, C. Riess, and M. Stamminger. Exploiting visual artifacts to expose deepfakes and face manipulations. In *IEEE Winter Conference on Applications of Computer Vision*, 2019.

[16] S. McCloskey and M. Albright. Detecting GAN-generated Imagery using Color Cues. *arXiv preprint arXiv:1812.08247v1*, 2018.

[17] H. Mo, B. Chen, and W. Luo. Fake Faces Identification via Convolutional Neural Network. In *Proc. of the 6th ACM Workshop on Information Hiding and Multimedia Security*, June 2018.

[18] National Institute of Standards and Technology. Media Forensics Challenge. <https://www.nist.gov/itl/iad/mig/media-forensicschallenge->, 2018.

[19] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved Techniques for Training GANs. In *Advances in Neural Information Processing Systems*, pages 2234–2242, 2016.

[20] S. Tariq, S. Lee, H. Kim, Y. Shin, and S. Woo. Detecting both machine and human created fake face images in the wild. In *Proc. of the 2nd International Workshop on Multimedia Privacy and Security*, pages 81–87, 2018.

[21] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner. Face2Face: Real-Time Face Capture and Reenactment of RGB Videos. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2387–2395, 2016.

[22] N. Yu, L. Davis, and M. Fritz. Attributing Fake Images to GANs: Analyzing Fingerprints in Generated Images. *arXiv preprint arXiv:1811.08180v1*, 2018.

[23] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision*, 2017.