Preliminary Forensics Analysis of DeepFake Images

Luca Guarnera
Department of Mathematics and Computer Science
University of Catania
Catania, Italy
luca.guarnera@unict.it

Oliver Giudice
Department of Mathematics and Computer Science
University of Catania
Catania, Italy
giudice@dmi.unict.it

Cristina Nastasi
Department of Mathematics and Computer Science
University of Catania
Catania, Italy
cristina.nastasi@unict.it

Sebastiano Battiato
Department of Mathematics and Computer Science
University of Catania
Catania, Italy
battiato@dmi.unict.it

Abstract—One of the most terrifying phenomenon nowadays is the Deepfake: the possibility to automatically replace a person's face in images and videos by exploiting algorithms based on deep learning. This paper will present a brief overview of technologies able to produce Deepfake images of faces. A forensics analysis of those images with standard methods will be presented: not surprisingly state of the art techniques are not completely able to detect the fakeness. To solve this, a preliminary idea on how to fight Deepfake images of faces will be presented by analysing anomalies in the frequency domain.

Index Terms—Deepfake, Multimedia Forensics, Generative Adversarial Networks

I. Introduction

Artificial intelligence technologies are evolving so rapidly that unthinkable new applications and services have emerged: one of them is the DeepFake. DeepFakes refers to all those multimedia contents synthetically altered or created by exploiting machine learning generative models. DeepFakes are image, audio or video contents that appear extremely realistic to humans specifically when they are used to generate and/or alter/swap image of faces. Various examples of DeepFake, involving celebrities, have already be seen on the internet: the insertion of Nicholas Cage[1] in movies where he did not act like "Fight Club" and "The Matrix" or the impressive video in which Jim Carrey[2] plays Shining in place of Jack Nicholson. Other more worrying examples are the video of Obama (Figure 1(a)), created by Buzzfeed[3] in collaboration with Monkeypaw Studios, or the video in which Mark Zuckerberg[4] (Figure 1(b)) claims a series of statements about the platform's ability to steal its users' data.

These Deepfakes have already been spread by mass media, also in Italy, where the satirical tv program "Striscia La Notizia[5], broadcasted in September 2019 a video of the ex-premier Matteo Renzi talking about his colleagues in a "not so respectful way (Figure 1 (c)). As we can imagine, DeepFakes may have serious repercussions on the veracity of the news

Fig. 1. Several examples of DeepFake: (a) Obama, created by Buzzfeed in collaboration with Monkeypaw Studios; (b) Mark Zuckerberg, created by artists Bill Posters and Daniel Howe in partnership with advertising company Canny; (c) Matteo Renzi, created by "Striscia la Notizia."

spread by the mass media while representing a new threat for politics, companies and personal privacy.

Deepfakes are evolving quickly and are becoming dangerous, not just for the reputation of the victims but also for security. In this dangerous scenario, tools are needed to unmask the deepfakes detect them or, at least, to mitigate the potential harm and abuse that can be done be means of these multimedia contents.

Several big companies, from Facebook to Microsoft, have decided to take action against this phenomenon: Google has created a database of fake videos [42] to support researchers who are developing new methods to detect them while Facebook and Microsoft have launched the Deepfake Detection Challenge initiative[6] which invites people from all over the world to create new tools to detect deepfakes and manipulated media.

The motivation of this paper is straightforward, after a brief overview of the state of the art in order to better understand what the technologies able to produce Deepfake are, a preliminary forensics analysis will be carried out. A first contribution to the field will be demonstrating that it is possible

---

[1] https://www.youtube.com/watch?v=-yQxsIWO2ic
[2] https://www.youtube.com/watch?v=Dx59bskG8dc
[3] https://www.youtube.com/watch?v=cQ54GDm1eL0
[4] https://www.youtube.com/watch?v=NbedWhzx1rs
[5] https://www.striscialanotizia.mediaset.it/video/il-fuorionda-di-matteo-renzi_59895.shtml

[6] https://deepfakedetectionchallenge.ai/

for the image forensics expert to find anomalies that could be related to how the Deepfake image is made. In fact standard image forensics tools are able to highlight some anomalies which the expert can analyze in deep to find specific anomalies in the frequency domain. In particular the anomalies, after Fourier transform, will be shown and will make clear that each kind of DeepFake creation technology has an easily detectable pattern. These evidence - as a preliminary result to the field - could lead to further and more sophisticated (even automatic) detection and analysis techniques.

The remaining part of the paper is organized as follow. Section II presents an overview of DeepFake creation technologies. Section III investigates how Image Forensics can fights DeepFakes. The proposed method is described in Section IV. Section V concludes the paper with the explanation of our future goal in this field.

II. An overview of Generative technologies

A. Brief introduction to GAN

Synthetic audiovisual media can be generated with a variety of techniques. Currently, the most popular of these techniques is the Generative Adversarial Network (GAN) due to its flexible applications and realistic outputs. Generative adversarial networks (GANs) were firstly introduced by Ian Goodfellow [19] in 2014. They propose a new framework for estimating generative models via an adversarial process, in which we simultaneously train two models: a generative model G that captures the data distribution, and a discriminative model D able to estimate the probability that a sample came from the training data rather than G. The training procedure for G is to maximize the probability of D making a mistake. This framework corresponds to a min-max two-player game. In the case of Deepfakes, the G can be thought as a team of counterfeiters trying to produce fake currency, while the D stands to the police, trying to detect the malicious activity. G and D can be made by any kind of generative model, in the original version they are implemented through deep neural networks.

There are many types of models proposed in the literature [25], [39], [50], [53]. Architecture variant GANs are mainly proposed for the purpose of different applications e.g., image completion [24], image super resolution [31], text-to-image generation [41] and image to image transfer [53]. The original GAN paper [19] employed fully-connected neural networks for both generator and discriminator. *Laplacian Pyramid of Adversarial Networks* is proposed for the production of higher resolution images from lower resolution input GAN [14]. *Deep Convolutional GAN* is the first work where a deconvolutional neural networks architecture [49] is applied. *Boundary Equilibrium GAN* uses an autoencoder architecture for the discriminator which was first proposed in [52]. *Progressive GAN* involves progressive steps toward the expansion of the network architecture [25]. This architecture uses the idea of progressive neural networks first proposed in [43]. *BigGAN* [8] has also achieved state-of-the-art performance on the ImageNet datasets.

B. Technologies for image creation

Focusing on DeepFake images of face, in [28] Lample et al. an "encoder-decoder architecture, (*Fader Networks*), able to generate different realistic versions of an input image by varying the values of the attributes was introduces: given an input image x with its attributes y, the encoder maps x to a latent representation z, and the decoder is trained to reconstruct x given (z, y). At inference time, a test image is encoded in the latent space and the user chooses the values of the attributes y that are sent to the decoder. A classifier learns how to predict the y attributes given the latent representation z during training. The encoder-decoder is trained so that the latent representation z must contain enough sufficients information to allow inputs reconstruction while the latent representation must prevent the classifier from predicting the correct attribute values. The authors have trained and tested Fader Network considering the CelebA dataset [33], obtaining a model that can significantly change the perceived value of the attributes while preserving the natural appearance of the input images. Tests have also been carried out on the Oxford-102 dataset [37] (containing about $9,000$ images of flowers classified in 102 categories), by changing the colour of the flower while keeping the background unchanged. Excellent results have also been achieved in this test. The code is available at https://github.com/facebookresearch/FaderNetworks.

In [44], Shen et al. a novel method based on *residual image learning* for face attribute manipulation is proposed. It can model the manipulation operation, as learning the residual image, defined as the difference between the original input image and the desired manipulated one. The proposed work focuses on the attribute-specific face area instead of the entire face which contains many redundant and irrelevant details. They develop a dual scheme able to learn two inverse attribute manipulations (one as the primal manipulation and the other as the dual manipulation) simultaneously. For each face attribute manipulation there are two image transformation networks called G_0 and G_1 and a discriminative network D. G_0 and G_1 that respectively simulate the primal and the dual manipulation. D classifies the reference images and generated images into three categories.

Several DeepFake based techniques that are present at the state of the art are limited regarding the management of more than two domains (for example, to change hair color, gender, age, and many others features in a face), since they should generate different models for each pair of image domains. A method capable of performing image-to-image translations on multiple domains using a single model has been proposed by Choi et al. [10] by means of a technology called *StarGAN*, a generative adversarial network. The main purpose was to define a scalable image-to-image translation model across multiple domains using a single generator and a discriminator. The authors used two different types of face datasets: CelebA [33] and RaFD dataset [29] Given a random label (for example hair color, facial expression, etc) as input, the network is able to perform an image-to-image translation operation considering

the given label. The results have been compared with other existing methods [32], [38], [53] and show how StarGAN manages to generate images of superior visual quality. The code is available at https://github.com/yunjey/stargan.

Wang et al. [48] propose Identity-Preserved Conditional Generative Adversarial Networks (*IPCGAN*), a framework for facial aging. IPCGAN is composed of three parts: a CGANs module (takes an input image and a target age to generate a new face with that age), an identity-preserved module (guarantees the aged face has the same input identity) and an age classifier (to ensure that the output has the desired age). The authors considered faces with different ages divided into 5 groups: 11-20, 21-30, 31-40, 41-50 and 50+. Given an image of the face x, they use the C_s information to indicate the age group to which x belongs. The aging of a face aims to generate a synthesized face of the target age group C_t. The framework has been trained and tested using the Cross-Age Celebrity Dataset (CACD) [9]. The performances of IPCGAN have been compared with acGAN [2] and CAAE [51] which achieve performance at vanguard for aging of the face. The qualitative and quantitative tests show that IPCGAN achieves the best results. Finally, IPCGAN can also be used to perform multi-attribute transfer tasks. The code is available at https://github.com/dawei6875797/Face-Aging-with-Identity-Preserved-Conditional-Generative-Adversarial-Networks.

The Style Generative Adversarial Network, or *Style-GAN* [26], proposes large changes to the generator model, including the use of a mapping network to map points in latent space to an intermediate latent space, to control style at each point in the generator model, and the introduction to noise as a source of variation at each point in the generator model. The resulting model is capable not only of generating impressively photorealistic high-quality photos of faces, but also offers control over the style of the generated image at different levels of detail through varying the style vectors and noise. In December 2018, the visual computing company NVIDIA, released an open source code for photorealistic face generation software created thanks to the StyleGAN algorithm [26]. Then, Uber's computer engineer Phillip Wang created the website https://thispersondoesnotexist.com/ and then on 11 February 2019 published it on the public group Facebook Artificial Intelligence and Deep Learning. StyleGAN algorithm is to be able to create realistic pseudo-portraits, difficult to judge as fakes. StyleGAN has also difficulty with the definition of the teeth and it cannot identify the backgrounds. Furthermore, there are often fluorescent spots, similar to water drops, which can appear anywhere on the image. To correct those imperfections in StyleGAN, Karras et al. made some improvements to the generator (including re-designed normalization, multi-resolution, and regularization methods) proposing StyleGAN2 [27].

III. FIGHTING DEEPFAKES WITH IMAGE FORENSICS

Image forgery and alteration is not a new problem introduced by DeepFakes. Counterfeiting an image with image editing tools like Photoshop is still very common as today

and the image/multimedia forensics science as already dealt with the problem [4]. There are several techniques that try to understand if a multimedia content is fake by means of various strategies to detect anomalies in the hidden structure of the multimedia content itself, exploiting noise, compression parameters, etc. Some try to reconstruct the history of the image [18] to identify the source-acquiring device or software, others instead analyze anomalies in the compressed JPEG domain [5], [17], [16] like Galvan et al. [16] who proposed a method which is able to recover the coefficients of the first compression process in a double compressed JPEG image to verify if there are altered elements. Another example from Battiato et al. [5] exploits the statistical distribution of the DCT coefficients in order to detect irregularities due to the presence of a signal overlapped on the original image. More recently, Giudice et al. [17] proposed a new analysis that can be carried out in the DCT domain able to automatically classify doubly compressed JPEG images with extremely high precision, giving forensics experts a tool to find the first evidence of image alteration. Not only the hidden structure of the image can be useful fakeness analysis. Even if the DeepFake images are extremely realistic, the visible contents could be analysed in order to find anomalies useful for detection.

In this context, Marra et al. [35] discussed the performance of various image-to-image translations detectors, both in ideal conditions and in the presence of compression, performed at the time of uploading to social networks. The study, conducted on a dataset of 36.302 images, shows that it is possible to obtain detection accuracy up to 95% both with conventional detectors and with deep learning based detectors, but only the latter continue to provide high accuracy, up to 89%, on compressed data.

Hsu et al. [23] also proposed an interesting method for detecting Deepfakes, called Deep Forgery Discriminator (DeepFD). Thanks to the implementation of a new discriminator that uses "contrastive loss it is possible to find the typical characteristics of the synthesized images generated by different GANs and therefore use a classifier to detect such fake images. For the training phase they used the CelebA dataset [33], considering 5 state-of-the-art GANs to generate the pool of false images [3], [22], [25], [34], [39]. Using the DeepFD they detected 94.7% of fake images generated by numerous state-of-the-art GANs, exceeding the other basic approaches present in the state of the art in terms of precision and recall rates. The code is available at https://github.com/jesse1029/Fake-Face-Images-Detection-Tensorflow.

As described by Guarnera et al. [20], the current GAN architectures that create Deepfake images, through convolution layers, leaves a fingerprint that characterizes that specific neural architecture. In order to capture this forensic trace, the authors used the Expectation-Maximization Algortihm [36] obtaining features able to distinguish real images from Deepfake ones.

Bonettini et al [7]. exploits Benford's law with the aim of identifying and analyzing statistical traces in order to detect if an image was generated by a generic GAN architecture. The
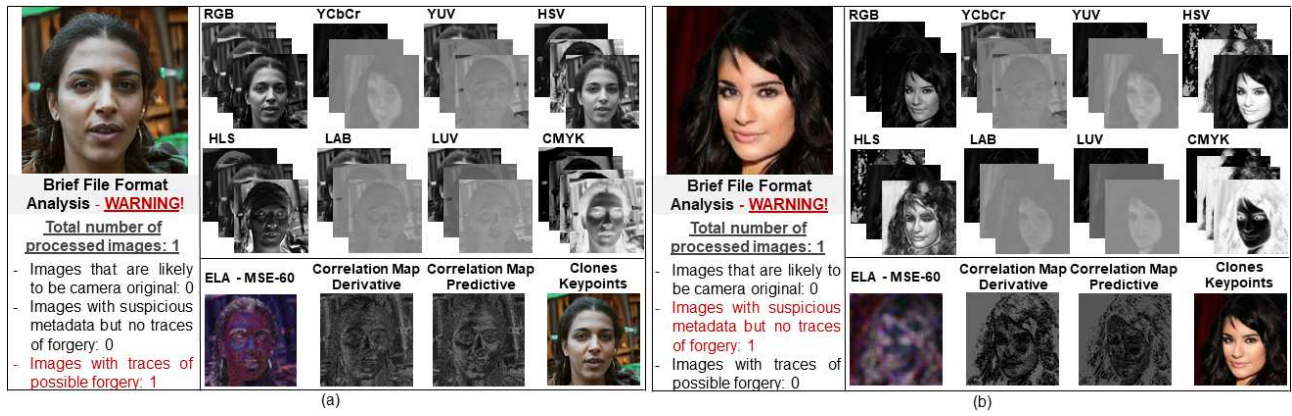
Fig. 2. Example of Analysis carried out with Amped Authenticate software. (a) Image generated by STYLEGAN. (b) Image generated by STARGAN. In both examples we show only some of the elements analyzed with Amped Authenticate, such as the analyzed different color spaces to understand if anomalies are found; ELA (Error Level Analysis) for Identification of spliced areas of the image that have been compressed differently; Correlation map to analyze and identify the correlation between the pixels of the image; Clones Keypoints to find parts of the image that appear to be cloned.

authors verify that Benford's law is not followed by GAN images and have proposed a number of related features to detect images generated by CycleGAN [53] and ProGAN [25]. From the conducted experiments, good detection results were obtained using a Random Forest classifier.

Wang et al. [47] investigated the possibility to create an universal detector able to identify the real images generated by a CNN regardless of the architecture or dataset used. They trained a classifier on a single CNN generator (ProGAN [25]). The conducted experiments demonstrated that this classifier detects synthesized images generated by different architectures and is also robust to JPEG compression, spatial blurring and scaling.

There are some specific techniques also for video which try to define if they are fakes. A video based DeepFake Detection method is described in the work of Gera et al. [21]. The authors used a CNN to extract frame-level features. In particular they use a Recurrent neural network (RNN) to train a classifier able to recognise if the video was manipulated or not by evaluating temporal inconsistencies introduced between the frames where face is modied. They evaluated the method with 600 videos: 300 fakes found on the web, 300 of real scenes taken from the HOHA dataset [30]. From experiments carried out by the authors, they get 97% accurate for fake detection.

Finally, Rossler et al. [42] proposed an automated benchmark for fake detection, based mainly on four manipulation methods: two computer graphics-based methods (Face2Face [46], FaceSwap[7]) and 2 learning-based approaches (Deepfakes[8], NeuralTextures [45]). The authors addressed the problem of fake detection as a binary classification for each frame of manipulated videos, considering different techniques present in the state of the art [1], [6], [11], [12], [15], [40].

A face tracking method is initially applied to the input image [46] (so as to work only on that particular region) to switch to the classification methods. The experiments con-

ducted show that XceptionNet achieves the best results. They also tested XceptionNet with images without face tracking information. In this case, however, the XceptionNet classifier has significantly lower accuracy. The code is available at https://github.com/ondyari/FaceForensics.

IV. DEEPFAKE FORENSICS ANALYSIS

In the previous paragraphs the most accurate techniques used nowadays to create DeepFake were described, and therefore different detection techniques were discussed. Most of the existing detection techniques are based on neural networks and it is very complicated to explain their results in a courtroom, giving their black-box features. Being able to understand the type of architecture used, the anomalous features and deterministically explaining why an image is a Deepfake is a process that is still almost completely absent in the literature.

In general, in the Multimedia Forensics [4] best practices are used to determine if an image is a fake, where, one tries to understand if it comes from a particular source, analyzing, information present in the metadata (how much present and not altered), the PNRU present in the images (fingerprint of the source), analysis of the coefficients obtained by JPEG compression, analysis in the Fourier domain and much more. If a specific information is not present, the suspicion arises that the data in question is not real. Nowadays, the current neural networks that generate DeepFake do not perform the same operations that are performed by any acquisition imaging source. This leads to obtain images with anomalies at the pixel level, obtaining a pattern that is not present if we consider an image generated by any device or software (camera, scanner, social network, etc.).

A forensics analysis was carried out on sample Deepfake Images by means of one of the most famous image forensics software "Amped Authenticate"[9]: it was employed to check if it is possible to identify whether an image, generated by a GAN, has anomalies, considering in particular the output of

[7]https://github.com/MarekKowalski/FaceSwap/

[8]https://github.com/deepfakes/faceswap/

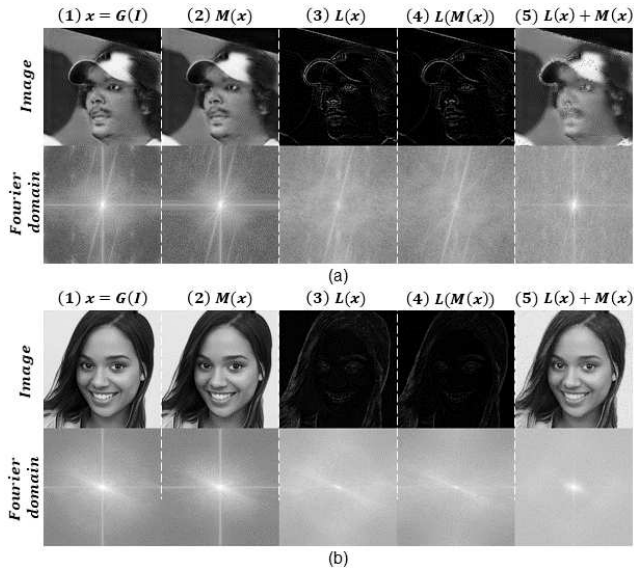[9]https://ampedsoftware.com/it/authenticate/

Fig. 3. Examples of analyzed images generated by (a) STARGAN and (b) STYLEGAN. Each image *I* of both datasets was converted to grayscale (1) and applied progressively: the Median filter (2), the Laplacian filter (3), the Laplacian filter (4) applied to the result of (2), the sum of the results between the Median and Laplacian filters (5). For each operation performed, we show the Fourier transform.

two types of technologies: StarGAN [10] and StyleGAN [26], briefly described Section II. In particular we analyzed JPEG structure of the image, we tried to infer Camera Identification (PNRU Identification), and then analysed those images in different color spaces (RGB, YCbCr, YUV, HSV, HLS, XYZ, LAB, LUV, CMYK); domains (ELA, DCT Map, JPEG Dimples Map, Blocking Artifacts, JPEG Ghosts Map, Fusion Map, Correlation Map, PRNU Map, PRNU Tampering, LGA) and by means of many forgery detection techniques (Clones Blocks, Clones Keypoints (Orb and Brisk). Figure 2 shows an overview of the results obtained and it is possible to see that in some cases the images may show specific anomalies while in other cases simple warnings are shown by the tool. However, this is not enough to define with certainty if the images are Deepfake, the only thing that could be inferred is that they are probably not-authentic and integrity is broken.

A deep analysis was then carried out in the frequency domain. Indeed, useful information can be obtained by working on Deepfake candidate" images after the Fourier transform. In general, a simple operation of Forgery or a Deepfake contains abnormal frequencies not present in real images. The application of convolutive filters with the respective Fourier spectra highlights the presence of a somewhat suspicious pattern, not present in real data. We performed several tests, using the Laplacian and median filters and the combination of them in order to enhance them (Figure 3). In the Fourier domain, as shown in Figure 3, it is possible to notice anomalous frequencies that substantially represent a pattern of that particular network used to generate fakes. This information is useful for identifying the type of neural network used and the areas in which that pattern is present. In Figure 3

the anomalies are clearly visible and are different for the two samples of each different techniques. Probably these patterns represents the way that deep neural network, and their convolutive layers, create the image, so they could be related to the hyper-parameters such as kernel masks employed in each generative technique. Obviously further investigation is needed but the conjunction of standard image forensics techniques that arises warning of "fakeness" and the detection of known anomalies (different for each technology) in the Fourier domain can achieve good results in terms of Deepfake Detection performance.

Forensics vs. anti-forensics is always an open game, and while we are dealing with detection methods, there are already a attempts seeking to hide those anomalies that were described above by introducing new camuoflage information into the fake images. This method was already proposed by Cozzolino et al. [13], in his work called "SpoC: Spoofing Camera Fingerprints in which they proposed a GAN-based approach capable of injecting traces of the camera fingerprint into images (thus probably reducing traces of the synthesis process), thus tricking avant-garde detectors.

V. CONCLUSION

In this paper we presented several techniques of creation and detection of the so-called Deepfakes and the related social and legal problems. It turns out to be very important to be able to create new methods that can counter this phenomenon. This could be done by analyzing details and traces of underlying generation process of the image (e.g. in the Fourier domain).

REFERENCES

[1] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen. Mesonet: a compact facial video forgery detection network. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–7. IEEE, 2018. 4

[2] G. Antipov, M. Baccouche, and J. Dugelay. Face aging with conditional generative adversarial networks. In *2017 IEEE international Conference on Image Processing (ICIP)*, pages 2089–2093. IEEE, 2017. 3

[3] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017. 3

[4] S. Battiato, O. Giudice, and A. Paratore. Multimedia forensics: discovering the history of multimedia contents. In *Proceedings of the 17th International Conference on Computer Systems and Technologies 2016*, pages 5–16. ACM, 2016. 3, 4

[5] S. Battiato and G. Messina. Digital forgery estimation into DCT domain: a critical analysis. In *Proceedings of the First ACM Workshop on Multimedia in Forensics*, pages 37–42, 2009. 3

[6] B. Bayar and M. C Stamm. A deep learning approach to universal image manipulation detection using a new convolutional layer. In *Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security*, pages 5–10, 2016. 4

[7] Nicolò Bonettini, Paolo Bestagini, Simone Milani, and Stefano Tubaro. On the use of benford's law to detect GAN-generated images. *arXiv preprint arXiv:2004.07682*, 2020. 3

[8] A. Brock, J. Donahue, and K. Simonyan. Large scale GAN training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 2

[9] B. Chen, C. Chen, and W. H Hsu. Cross-age reference coding for age-invariant face recognition and retrieval. In *European Conference on Computer Vision*, pages 768–783. Springer, 2014. 3

[10] Y. Choi, M. Choi, M. Kim, J. Ha, S. Kim, and J. Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8789–8797, 2018. 2, 5

[11] F. Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1251–1258, 2017. 4

[12] D. Cozzolino, G. Poggi, and L. Verdoliva. Recasting residual-based local descriptors as convolutional neural networks: an application to image forgery detection. In *Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security*, pages 159–164, 2017. 4

[13] D. Cozzolino, J. Thies, A. Rössler, M. Nießner, and L. Verdoliva. Spoc: Spoofing Camera Fingerprints. *arXiv preprint arXiv:1911.12069*, 2019. 5

[14] E. L Denton, S. Chintala, R. Fergus, et al. Deep generative image models using a Laplacian pyramid of adversarial networks. In *Advances in Neural Information Processing Systems*, pages 1486–1494, 2015. 2

[15] J. Fridrich and J. Kodovsky. Rich models for steganalysis of digital images. *IEEE Transactions on Information Forensics and Security*, 7(3):868–882, 2012. 4

[16] F. Galvan, G. Puglisi, A. Bruna, and S. Battiato. First quantization coefficient extraction from double compressed jpeg images. In *International Conference on Image Analysis and Processing*, pages 783–792. Springer, 2013. 3

[17] O. Giudice, F. Guarnera, A. Paratore, and S. Battiato. 1-D DCT domain analysis for JPEG double compression detection. In *Proceeedings of International Conference on Image Analysis and Processing*, pages 716–726. Springer, 2019. 3

[18] O. Giudice, A. Paratore, M. Moltisanti, and S. Battiato. A classification engine for image ballistics of social data. In *Proceeedings of International Conference on Image Analysis and Processing*, pages 625–636. Springer, 2017. 3

[19] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014. 2

[20] L. Guarnera, O. Giudice, and S. Battiato. Deepfake detection by analyzing convolutional traces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 666–667, 2020. 3

[21] D. Güera and E. J Delp. Deepfake video detection using recurrent neural networks. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6. IEEE, 2018. 4

[22] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C Courville. Improved training of wasserstein gans. In *Advances in neural information processing systems*, pages 5767–5777, 2017. 3

[23] C. Hsu, C. Lee, and Y. Zhuang. Learning to detect fake face images in the wild. In *2018 International Symposium on Computer, Consumer and Control (IS3C)*, pages 388–391. IEEE, 2018. 3

[24] S. Iizuka, E. Simo-Serra, and H. Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics (ToG)*, 36(4):1–14, 2017. 2

[25] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 2, 3, 4

[26] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 3, 5

[27] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. Analyzing and improving the image quality of stylegan. *arXiv preprint arXiv:1912.04958*, 2019. 3

[28] G. Lample, N. Zeghidour, N. Usunier, A. Bordes, L. Denoyer, and M. Ranzato. Fader networks: Manipulating images by sliding attributes. In *Advances in Neural Information Processing Systems*, pages 5967–5976, 2017. 2

[29] O. Langner, R. Dotsch, G. Bijlstra, D. HJ Wigboldus, S. T Hawk, and AD Van Knippenberg. Presentation and validation of the radboud faces database. *Cognition and emotion*, 24(8):1377–1388, 2010. 2

[30] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008. 4

[31] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4681–4690, 2017. 2

[32] M. Li, W. Zuo, and D. Zhang. Deep identity-aware transfer of facial attributes. *arXiv preprint arXiv:1610.05586*, 2016. 3

[33] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3730–3738, 2015. 2, 3

[34] X. Mao, Q. Li, H. Xie, R. YK Lau, Z. Wang, and S. Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2794–2802, 2017. 3

[35] F. Marra, D. Gragnaniello, D. Cozzolino, and L. Verdoliva. Detection of gan-generated fake images over social networks. In *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 384–389. IEEE, 2018. 3

[36] T. K Moon. The expectation-maximization algorithm. *IEEE Signal Processing Magazine*, 13(6):47–60, 1996. 3

[37] M. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729. IEEE, 2008. 2

[38] G. Perarnau, J. Van De Weijer, B. Raducanu, and J. M Álvarez. Invertible conditional GANs for image editing. *arXiv preprint arXiv:1611.06355*, 2016. 3

[39] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 2, 3

[40] N. Rahmouni, V. Nozick, J. Yamagishi, and I. Echizen. Distinguishing computer graphics from natural images using convolution neural networks. In *2017 IEEE Workshop on Information Forensics and Security (WIFS)*, pages 1–6. IEEE, 2017. 4

[41] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*, 2016. 2

[42] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1–11, 2019. 1, 4

[43] A. A Rusu, Neil C R., G. Desjardins, H. Soyer, J. Kirkpatrick, K. Kavukcuoglu, R. Pascanu, and R. Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016. 2

[44] W. Shen and R. Liu. Learning residual images for face attribute manipulation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4030–4038, 2017. 2

[45] J. Thies, M. Zollhöfer, and M. Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019. 4

[46] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner. Face2face: Real-time face capture and reenactment of RGB videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2387–2395, 2016. 4

[47] S. Wang, O. Wang, R. Zhang, A. Owens, and A. Efros. Cnn-generated images are surprisingly easy to spot...for now. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,*, 2020. 4

[48] Z. Wang, X. Tang, W. Luo, and S. Gao. Face aging with identity-preserved conditional generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7939–7947, 2018. 3

[49] M. D Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, pages 818–833. Springer, 2014. 2

[50] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5907–5915, 2017. 2

[51] Z. Zhang, Y. Song, and H. Qi. Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5810–5818, 2017. 3

[52] J. Zhao, M. Mathieu, and Y. LeCun. Energy-based generative adversarial network. *arXiv preprint arXiv:1609.03126*, 2016. 2

[53] J. Zhu, T. Park, P. Isola, and A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2223–2232, 2017. 2, 3, 4