Evasion Attack on Deepfake Detection via DCT Trace Manipulation

Luca Guarnera[1 [0000−0001−8315−351X]], Francesco Guarnera[1 [0000−0002−7703−3367]], Alessandro Ortis[1 [0000−0003−3461−4679]], Sebastiano Battiato[1 [0000−0001−6127−2470]], and Giovanni Puglisi[2 [0000−0001−6222−9933]]

[1] Department of Mathematics and Computer Science University of Catania, Italy
{luca.guarnera,francesco.guarnera,alessandro.ortis,sebastiano.battiato}@unict.it
[2] Department of Mathematics and Computer Science University of Cagliari, Italy
puglisi@unica.it

Abstract. In the last years, lots of approaches devoted to recognize fake images have been developed. Some of them, exploiting traces left in the frequency domain by the fake image generators, were able to achieve satisfactory results also employing simple classifiers. In this paper, a novel white-box evasion attack was introduced to deceive a specific class of frequency-based deepfake detectors exploiting DCT (Discrete Cosine Transform) features. Specifically, statistics computed from the distribution of the AC frequencies computed from fake images are aligned to the corresponding values extracted from authentic images. The robustness of both classical and state-of-the-art DCT-based classifiers has been tested with respect to the proposed attack considering fake images generated by Generative Adversarial Networks and Diffusion Models.

Keywords: Adversarial imaging · Withe-box attack · DCT analysis · Fake images · Adversarial discriminator .

1 Introduction

Deepfakes represent one of the most complex and concerning challenges within the contexts of cybersecurity and digital ethics. Crafted through the exploitation of sophisticated machine learning methodologies, such as Generative Adversarial Networks (GANs) [16] and Diffusion Models (DMs) [19], deepfakes have the capability of generating realistic and convincing audiovisual content, designed to manipulate or deceive human viewers. However, these generative models leave behind identifiable traces in deepfake images which can be effectively detected by deepfake detectors. These traces often manifest as clear patterns in frequency domain and in particular in the DCT one [6, 8, 14, 29]. By leveraging these distinctive features, a classifier can be trained to discriminate between deepfake and real images by pinpointing proper traces or using specific combinations of statistics computed from AC coefficient distributions. It is feasible to detect unique AC coefficient statistics associated with particular generative models,

serving then as distinctive fingerprints, effectively imprinting its signature onto the generated images.

This paper presents an innovative white-box evasion attack strategy aimed at deceiving deepfake detection systems which leverage on the analysis of Discrete Cosine Transform (DCT) traces. Our approach involves refining synthetic images by precisely manipulating the AC coefficients associated with generative models. Specifically, taking inspiration from adversarial attack approaches [17], given a synthetic image F, the proposed methodology introduces carefully crafted perturbations to the statistics computed in the DCT domain to align them to the ones extracted from genuine images. Once the statistics alignment has been performed employing the histogram matching method, the inverse DCT transform is applied to generated a resilient synthetic image able to deceive the deepfake detection mechanisms. This strategy is grounded in the understanding (white-box evasion) that deepfake detection systems often rely on discrepancies in statistical properties between genuine and synthetic images, such as differences in DCT coefficients' distributions. By properly modifying these coefficients to match those of genuine images, our approach aims to overcome the detection of fake images, causing them to be classified as real. For sake of generalization, both GAN and DM AI-engines have been exploited. The effectiveness of our method is supported by empirical evidences and experiments, which demonstrated the ability of the method to fool deepfake detection algorithms. Our findings contribute to the ongoing talks surrounding the arms race between deepfake generation and detection technologies, highlighting the need for robust countermeasures to mitigate the proliferation of misleading media content.

The paper is structured as follows: depth analysis of the state-of-the-art in the field of deepfakes detection is presented in Section 2. Section 3 describes the proposed solution whereas Section 4 introduces the dataset used in the experimental phase. The experimental results are detailed and analysed in Section 5. We conclude the paper by discussing the implications of the results obtained and suggesting potential hints for future investigation and research in this area.

2 Related Works

Adversarial Machine Learning (AML) encompasses the study of vulnerabilities in machine learning models and systems, particularly in the context of adversarial attacks and defenses. AML techniques have been instrumental in understanding and mitigating vulnerabilities associated with predictive methods [17]. Scientific efforts delving into AML, including the creation of adversarial examples and methods aimed at evading AI models, offer invaluable insights into the complexity of these systems. Such insights not only enhance our comprehension of the underlying mechanisms but also pave the way for fortifying the robustness and trustworthiness of AI systems against emerging threats like deepfakes.

In the context of deepfake detection, DCT-based methods exploit traces left by the generative models in the DCT domain, which can be used to differentiate between authentic and manipulated images. Lewis et al. [8] illustrated the utilization of DCT spectrum to extract spectral information from individual

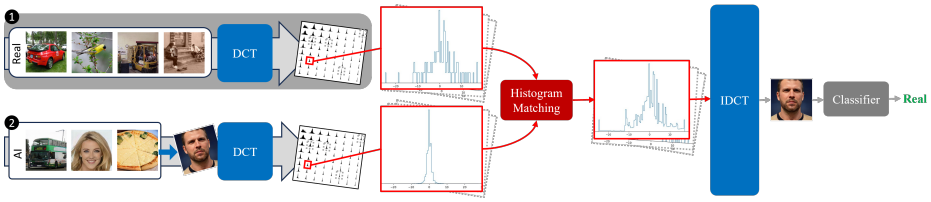


Fig. 1: Proposed pipeline. First (1), the overall statistics (in mean) of the real dataset are extracted using DCT. Then, given a deepfake image as input (2), the DCT is applied followed by the histogram matching algorithm. Finally, the adversarial deepfake image is reconstructed through IDCT. The latter is given as input to the machine learning classifiers with the aim of obtaining a misclassification, i.e., real.

frames, thereby improving deepfake detection. The investigation by Frank et al. [12] found that frequency representation surpasses state-of-the-art techniques in detecting deep fake images automatically. Recent studies have proved the effectiveness of frequency domain-based techniques in identifying distinct anomalous fingerprints associated with deepfakes, yielding notable outcomes. Of particular interest are deepfake detection methods leveraging the DCT. These methods encompass direct application of DCT to images, as demonstrated by Joel et al. [13], as well as extraction of features from DCT blocks, akin to JPEG compression, as showcased by Giudice et al. [14]. Both approaches have demonstrated substantial effectiveness in detecting and characterizing the unique digital signatures left by the generative architectures employed in deepfake creation. More recently in [29] a proper analysis of the DCT traces in the generative AI-domain has been exploited by considering several architectures, including DMs. This methodology represents a significant advancement in both understanding and detecting deepfakes. Recently, several adversarial methods have been proposed in literature aimed at fooling deepfake detection classifiers exploiting Fourier spectrum discrepancies. The approach outlined in [10] identified the up-sampling layers within generative models as a key factor contributing to these models' inability to accurately replicate the spectral distributions of authentic images. These up-sampling layers often introduce significant spectral distortion into the generated images. Building on this observation, the authors showcased a method for reliably detecting deepfakes, regardless of the underlying architecture. Furthermore, the authors suggest incorporating a spectral regularization term into the attack optimization process to mitigate this issue. The work in [3] delved into the frequency disparities observed in fake images, demonstrating that these disparities can be mitigated through minor changes in the final upsampling step of the generative model. Additionally, the researchers produced counterexample images capable of bypassing a forensic detection detector based on Fourier spectrum attributes. The authors of [9] have proposed two methods to mitigate the intensity of patterns resulting from spectral artifacts. Firstly, they introduced a CycleGAN capable of mapping images from the domain of fake images con-

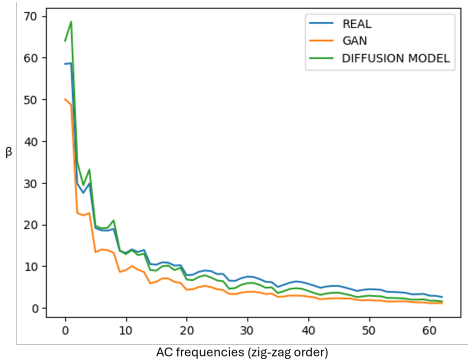


Fig. 2: Average distributions of β for each class. AC coefficients represented following classic “zig-zag" ordering.

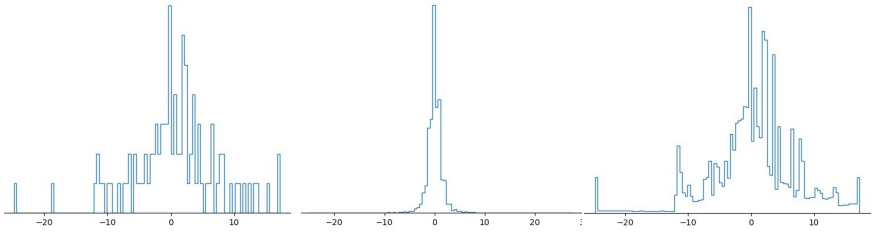
taining artifacts to the domain of real images without artifacts (SpectralGAN). Secondly, they proposed a method that incorporates the disparities in spectrum between the two domains by simply subtracting the average spectrum differences between fake and real images (Spectrum Difference Normalization). Both methods are then combined with a dictionary-based approach to correct discrepancies in the spectral power distribution (Power Distribution Correction - PDC) within GAN-generated fake images. The findings from [9] suggest that detectors relying on the analysis of spectral characteristics of fake images may not be resilient in detecting GAN-generated fakes. Indeed, artifacts within such spectra can be easily mitigated by malicious attackers, resulting in visually similar images post-enhancement.

3 Proposed Method

Synthetic multimedia content created through AI technologies (GAN, DM) allows the generation of realistic images. Although the detection of these artificial images is visually difficult, many classifiers have been proposed in recent years that can distinguish real from fake images. As described in the previous Section, a class of these approaches exploits anomalies in the frequency domain introduced by the generator, achieving a good level of discrimination. To test the robustness of these classifiers, an adversarial attack in the frequency domain has been designed. Specifically, the proposed solution aims to bring the discriminant frequencies of fake images closer to the real ones, confusing the classifier during the decision process. In particular, we paid attention to the features related to the Discrete Cosine Transform (DCT) [8, 11, 14, 29]. Figure 1 summarizes the proposed pipeline.

3.1 Extraction and analysis of β coefficients

DCT is a Fourier-related transform widely exploited in signal processing and data compression. Such as example, the well-known JPEG compression algorithm [32], after partitioning the input image into 8×8 blocks applies the aforementioned transformation. The values of the DCT coefficients can be regarded



(a) h_i^r (b) $h_i(F)$ (c) $h_i^r(F)$

Fig. 3: Histogram matching $h_i^r(F)$ (c) between $h_i(F)$ (a) and h_i^r (b)

as the relative amount of the two-dimensional spatial frequencies contained in the 64 input points. The first of these coefficients (top left, position $(0, 0)$) called DC, represents the average brightness level of the entire 8×8 block, whereas the other ones, called AC, provide a detailed picture of the brightness variations within the block. Lam et al. [24] demonstrated that AC coefficients can be effectively modelled by means of a zero-centered Laplacian distribution:

$$P(x|\beta) = \frac{1}{2\beta} \exp\left(-\frac{|x|}{\beta}\right) \tag{1}$$

where β is the scale parameter estimated as $\sigma/\sqrt{2}$ and σ corresponds to the standard deviation of the AC coefficient distributions. Taking inspiration from the works [14] and [11], our research focuses specifically on β parameters computed from AC distributions, which showed significant discriminating properties between real digital images and those generated through generative models.

3.2 Features manipulation

The proposed methodology consists of identifying and manipulating specific frequency bands in the synthetic images in order to make them similar to the frequency statistics of real data. To do this, the images were analyzed in the frequency domain using the Discrete Cosine Transform. Specifically, an input fake image F generated by a deepfake generative model G_i, is divided into non-overlapped blocks (f) of size 8×8 pixels. At each block, the DCT is applied as follows:

$$DCT(u, v) = \frac{C(u) \cdot C(v)}{4} \sum_{x=0}^{7} \sum_{y=0}^{7} f(x, y) \cdot t(x, u) \cdot t(y, v) \tag{2}$$

with $C(p) = \begin{cases} \frac{1}{\sqrt{2}} & \text{if } p = 0 \\ 1 & \text{otherwise} \end{cases}$, $t(z, \epsilon) = \cos\left[\frac{(2z+1)\epsilon\pi}{16}\right]$ and $f(x, y)$ the pixel intensity at coordinates (x, y) in the image block.

In order to design the proposed attack solution, the average values of β computed from the AC frequency distributions have been analysed. Specifically, three classes have been considered: real and deepfake images generated by GANs and

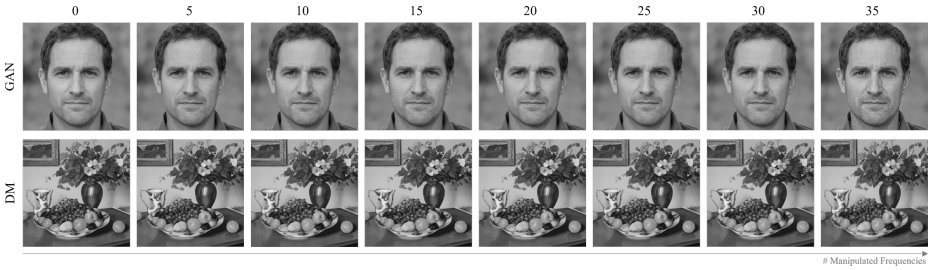


Fig. 4: Examples of images at varying numbers of manipulated frequencies.

Diffusion Models (see Section 4 for further details about the employed dataset). As can be seen from Figure 2, average β values related to the three classes, although sharing a similar decreasing pattern, are quite discriminative and have been exploited to design effective fake image detectors.

Given then a fake image F with beta values $\beta_i(F)$ $(i \in 1, 2, \ldots, 63)$ and knowing the average β values of real images β_i^r the main idea of the proposed attack is to perform a proper perturbation of input image to move its $\beta_i(F)$ to β_i^r deceiving then the classifiers based on these DCT statistics. It is easy to understand that the choice of AC frequencies to perturb is relevant for the generation of the output image. The authors of [29] carried out analyses to understand the most discriminative ones training a proper classifier and selecting just 35 AC frequencies to generate the corresponding adversarial images. Given the fake input image F the proposed approach can be summarized as follows:

1. image partitioning into 8×8 blocks and DCT application on each block;
2. extraction of the histograms $h_i(F)$ for each AC frequency to be perturbed;
3. generation of ideal histogram h_i^r from β_i^r values related to real distributions (only for AC frequencies to be modified);
4. generation of perturbed histogram $h_i^r(F)$ through histogram matching [15] between $h_i(F)$ and h_i^r (example is shown in Figure 3);
5. reconstruction of 8×8 blocks after perturbation;
6. inverse DCT (IDCT) on each 8×8 block with truncation on range $0 - 255$ to reconstruct the perturbed image.

4 Dataset Details

The dataset employed to properly evaluate the proposed solution was built starting from a subset of [29]: 4,449 images generated by different GAN architectures, and 1,449 images produced by Diffusion Models. Although, imbalance between classes, tests are conducted separately without involving any machine learning solution in the generation of the adversarial image (see Section 5). As regards real data we collected a total of 40.000 images from CelebA [26], FFHQ [22], and other different. sources [25, 7] (10.000 for each source). In accordance with the research goals, we prioritized the diversity of the inputs in order to guarantee that the dataset adequately represents a range of image generation techniques,

Type	Architecture Name	# Image Used
GAN GENERATED	GauGAN [28]	4000
	BigGAN [2]	2600
	ProGAN [20]	1000
	StarGAN [5]	6848
	AttGAN [18]	6005
	GDWCT [4]	3367
	CycleGAN [33]	1047
	StyleGAN [22]	4705
	StyleGAN2 [23]	7000
	StyleGAN3 [21]	1000
DM GENERATED	DALL-E 2 [30]	3423
	DALL-E MINI	1000
	Glide [27]	2000
	Latent Diffusion [31]	4000
	Stable Diffusion	5000

Table 1: Number of images used per category (GAN, DM).

regardless of their individual content. Note that the generation of the adversarial images was done as described in Section 3, working only on the luminance channel. For this reason the adversarial dataset is composed of grayscale images (see example in Figure 4). Each of the aforementioned image (4,449 from GAN and 1,449 from DM) has been perturbed employing 7 different sets of β_i^r generating a total of 41,286 grayscale adversarial images (31,143 from GAN and 10,143 from DM). Specifically, different attacks have been conducted modifying an increasing number of AC frequencies (i.e., 5, 10, 15, 20, 25, 30, 35) starting from the highest to the lowest ones. Table 1 summarizes the number of collected images[3].

5 Experimental Results

As described in Section 3, the proposed solution modifies the fake image F in the frequency domain in order to make it similar, in terms of DCT statistics, to the corresponding ones of the real pictures. However, it is evident that carry out too strong perturbations on DCT distributions could cause artifact visible for human eyes. In order to validate our method a series of experiments has been then performed. In particular, some variables were introduced in the tests as follows:

[3] Stable Diffusion and DALL-E MINI were generated from https://github.com/CompVis/stable-diffusion; https://github.com/borisdayma/dalle-mini.

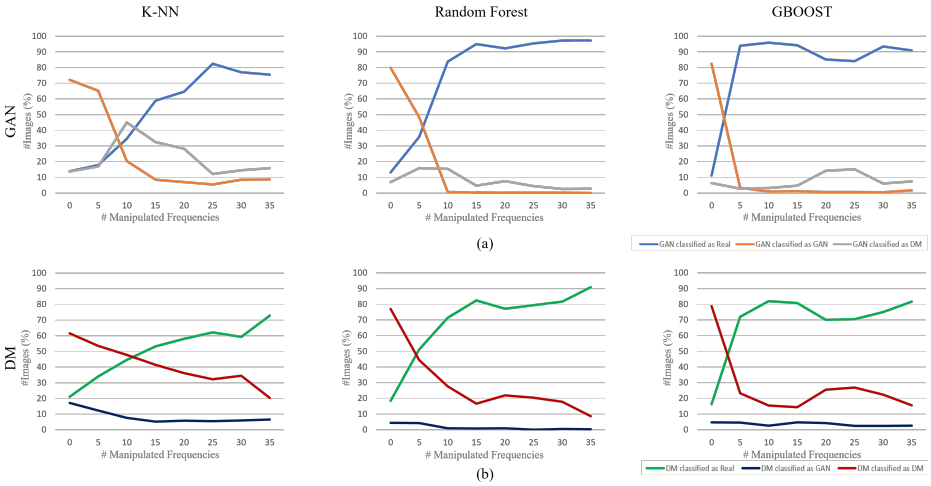


Fig. 5: Experimental results of the proposed method on images generated by GAN (a) and DM (b). Different columns are referred to the specific employed classifiers. For each plot, the percentage of images classified as real, images classified correctly, and images classified as the opposite generative model technology are given.

- The generative models: the experiments were performed considering both GAN and DM in order to understand the different behavior of different classes of generators;
- The modified frequencies: considering the set of 35 AC frequencies selected in [29], the proposed method was tested modifying an increasing number of frequencies from the highest to the lowest ones (from 5 to 35 at step of 5), to understand the right trade-off between the quality of the visual output and the effectiveness of the attack;
- The classifiers: 3 different classifiers based on Machine Learning approaches (K-NN, GBoost and Random Forest) were tested on adversarial images generated by our method.

Some examples of modified images are reported in Figure 4 whereas results achieved by the proposed attack considering three classifiers have been presented in Figure 5. For each plot, classification performances obtained with the original deepfake images (label 0) as input are compared with those achieved taking into account the modified pictures at increasing of the number of involved frequencies $\{5, 10, 15, \ldots, 35\}$. Specifically, the number of deepfake images classified as real, the number of images classified correctly, and the number of images classified as the opposite generative model technology (GAN or DM) are highlighted in each plot. In order to perform a peculiar and accurate analysis on the performance of these attacks, two different graphs were created with respect to the type of generative model (GAN and DM).

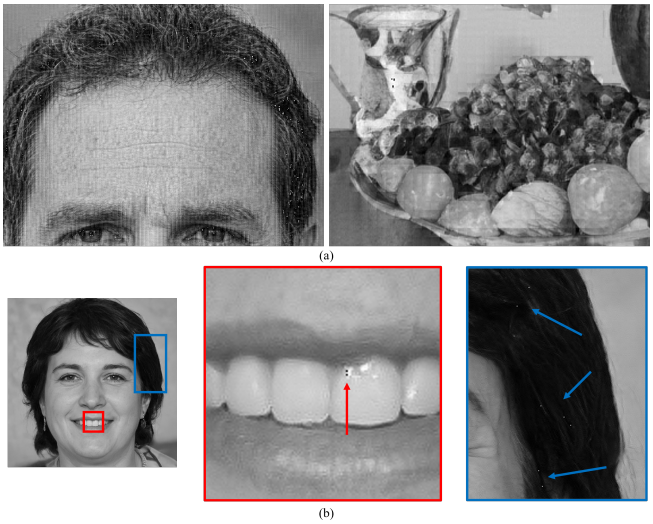


Fig. 6: (a) Images of Figure 4 with 35 manipulated frequencies. Artifacts in the form of blocking are present. (b) Image with 20 manipulated frequencies, in which the attack introduced salt-and-pepper noise in the teeth and hair.

From the obtained results, it can be observed that, in general, the performance of all classifiers collapses, even when a minimum number of frequencies is changed. It can be seen, however, that DM images seem to be more resistant to these types of attacks, as the classification performance decreases more linearly than GAN-generated images. Moreover, the most robust classifier turns out to be K-NN. Thus, it is very likely that deepfake detectors based on more sophisticated metric learning approaches may be more robust to this type of attack. The Random Forest and GBoost classifiers are based on well-defined "rules and weights" and, therefore, turn out to be less robust. In the latter scenario, we get very good results in solving the task at hand.

The quality of the attacked images plays a key role in adversarial attacks, as image quality must be preserved despite the attack. The proposed approach succeeds in fooling classifiers, but from a visual and perceptual perspective, the more frequencies are attacked, and the more artifacts are introduced into the images (Figure 6). It can be observed that as the number of frequencies attacked increases, the blocking effect (due to the nature of the proposed pipeline), is emphasized. In other cases, such as the one shown in Figure 6 (b), even though minimally and not so visually perceptible, the images after the attack of 20 frequencies contain salt-and-pepper noise. In order to investigate, from a quantitative point of view, the structural difference between the attacked and original images, the SSIM and PSNR metrics were calculated (averaged) on the GAN and DM sets separately. The SSIM metric will return a value that will tend to -1 in the case where the images are very different from each other, 1 on the other hand in the case where the compared images are very similar. The higher the PSNR metric, the greater the "similarity" to the original image.

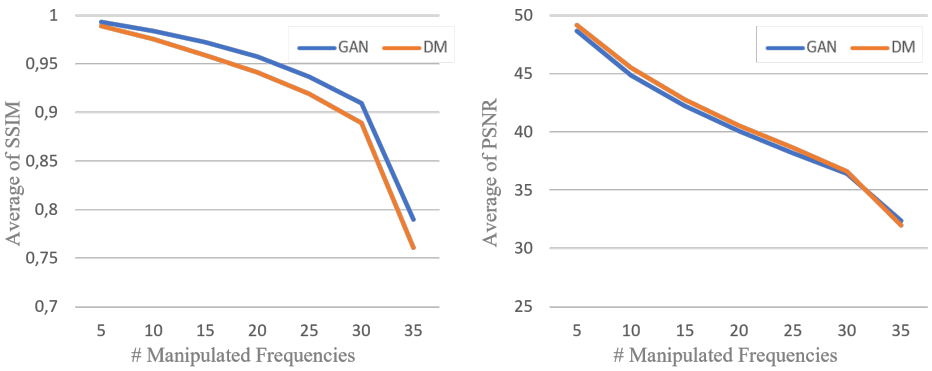


Fig. 7: (a) Mean of SSIM and (b) mean of PSNR calculated for all involved images as the number of manipulated frequencies varies.

From Figure 7, it can be deduced that by attacking a maximum of 15 frequencies, the image not only contains fewer artifacts, but still tends to be very close to the original, and the classifiers (especially Random Forest and GBoost) fail to correctly classify the data under analysis.

6 Conclusion and Future Works

In this work, we perform an evasion attack on DCT-based deepfake detection systems by generating perturbations that are imperceptible to the human eye but significantly alter the DCT coefficients of the image. These perturbations cause misclassification or evasion of the detection system, leading to the acceptance of deepfake images as authentic. Defense mechanisms against adversarial attacks in DCT-based deepfake detection often involve robust feature extraction and classifier training. By incorporating robust feature extraction techniques that are less susceptible to adversarial perturbations, and training classifiers on diverse datasets that include adversarial examples, it is possible to improve the resilience of detection systems against adversarial attacks [1]. This paper provides useful insights to further investigate the interplay between adversarial machine learning and DCT-based methods, which is pivotal for developing more robust deepfake detection systems capable to prevent adversarial attacks. Indeed, the proposed evasion method not only exposes vulnerabilities but also provides insights for defining more resilient AI models that can effectively counter such sophisticated attacks. Consequently, ongoing research in this area aims to further enhance the effectiveness and reliability of deepfake detection methods in the face of evolving adversarial threats.

7 Acknowledgment

This study has been partially supported by SERICS (PE00000014) under the MUR National Recovery and Resilience Plan funded by the European Union - NextGenerationEU.

References

1. Bai, T., Luo, J., Zhao, J., Wen, B., Wang, Q.: Recent advances in adversarial training for adversarial robustness. arXiv preprint arXiv:2102.01356 (2021)
2. Brock, A., Donahue, J., Simonyan, K.: Large scale GAN training for high fidelity natural image synthesis. In: International Conference on Learning Representations (2018)
3. Chandrasegaran, K., Tran, N., Cheung, N.: A closer look at Fourier spectrum discrepancies for CNN-generated images detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7200–7209 (2021)
4. Cho, W., Choi, S., Park, D.K., Shin, I., Choo, J.: Image-to-image translation via group-wise deep whitening-and-coloring transformation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10639–10647 (2019)
5. Choi, Y., Choi, M., Kim, M., Ha, J., Kim, S., Choo, J.: StarGAN: unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8789–8797 (2018)
6. Concas, S., Perelli, G., Marcialis, G.L., Puglisi, G.: Tensor-based deepfake detection in scaled and compressed images. In: 2022 IEEE International Conference on Image Processing (ICIP). pp. 3121–3125 (2022)
7. Corvi, R., Cozzolino, D., Zingarini, G., Poggi, G., Nagano, K., Verdoliva, L.: On the detection of synthetic images generated by diffusion models. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023)
8. Das, A.K., Mukhopadhyay, S., Dalui, A., Bhattacharya, R., Naskar, R.: Fighting deepfakes by detecting DCT frequency anomalies. In: 2023 International Symposium on Devices, Circuits and Systems (ISDCS). vol. 1, pp. 1–5. IEEE (2023)
9. Dong, C., Kumar, A., Liu, E.: Think twice before detecting GAN-generated fake images from their spectral domain imprints. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7865–7874 (2022)
10. Durall, R., Keuper, M., Keuper, J.: Watch your up-convolution: CNN based generative deep neural networks are failing to reproduce spectral distributions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7890–7899 (2020)
11. Farinella, G., Ravì, D., Tomaselli, V., Guarnera, M., Battiato, S.: Representing scenes for real-time context classification on mobile devices. Pattern Recognition **48**(4), 1086–1100 (2015)
12. Frank, J., Eisenhofer, T., Schönherr, L., Fischer, A., Kolossa, D., Holz, T.: Leveraging frequency analysis for deep fake image recognition. In: International Conference on Machine Learning. pp. 3247–3258. PMLR (2020)

13. Frank, J., Eisenhofer, T., Schönherr, L., Fischer, A., Kolossa, D., Holz, T.: Leveraging frequency analysis for deep fake image recognition. In: Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event. vol. 119, pp. 3247–3258. PMLR (2020)
14. Giudice, O., Guarnera, L., Battiato, S.: Fighting deepfakes by detecting GAN DCT anomalies. Journal of Imaging **7**(8), 128 (2021)
15. Gonzales, R.C., Woods, R.E.: Digital image processing. Pearson India (January 1, 2019) (2019)
16. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in Neural Information Processing Systems **27** (2014)
17. Goodfellow, I., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
18. He, Z., Zuo, W., Kan, M., Shan, S., Chen, X.: AttGAN: facial attribute editing by only changing what you want. IEEE Transactions on Image Processing **28**(11), 5464–5478 (2019)
19. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems **33**, 6840–6851 (2020)
20. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of GANs for improved quality, stability, and variation. In: International Conference on Learning Representations (2018)
21. Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., Aila, T.: Alias-free generative adversarial networks. Advances in Neural Information Processing Systems **34**, 852–863 (2021)
22. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4401–4410 (2019)
23. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of StyleGAN. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8110–8119 (2020)
24. Lam, E.Y., Goodman, J.: A mathematical analysis of the DCT coefficient distributions for images. IEEE Transactions on Image Processing **9**(10), 1661–1666 (2000)
25. Leotta, R., Giudice, O., Guarnera, L., Battiato, S.: Not with my name! Inferring artists' names of input strings employed by diffusion models. In: International Conference on Image Analysis and Processing. pp. 364–375. Springer (2023)
26. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep Learning Face Attributes in the Wild. In: Proceedings of International Conference on Computer Vision (ICCV) (December 2015)
27. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. In: International Conference on Machine Learning. pp. 16784–16804. PMLR (2022)
28. Park, T., Liu, M., Wang, T., Zhu, J.: GauGAN: semantic image synthesis with spatially adaptive normalization. In: ACM SIGGRAPH 2019 Real-Time Live! pp. 1–1 (2019)
29. Pontorno, O., Guarnera, L., Battiato, S.: On the exploitation of DCT-traces in the generative-AI domain. arXiv preprint arXiv:2402.02209 (2024)
30. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with CLIP latents. arXiv preprint:2204.06125 **1**(2), 3 (2022)

31. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10684–10695 (2022)
32. Wallace, G.K.: The JPEG still picture compression standard. Communications of the ACM **34**(4), 30–44 (1991)
33. Zhu, J., Park, T., Isola, P., Efros, A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2223–2232 (2017)