Misleading Deep-Fake Detection with GAN Fingerprints

Vera Wesselkamp*, Konrad Rieck*, Daniel Arp† and Erwin Quiring*
* *Technische Universität Braunschweig, Germany*
† *Technische Universität Berlin, Germany*

***Abstract*—Generative adversarial networks (GANs) have made remarkable progress in synthesizing realistic-looking images that effectively outsmart even humans. Although several detection methods can recognize these deep fakes by checking for image artifacts from the generation process, multiple counterattacks have demonstrated their limitations. These attacks, however, still require certain conditions to hold, such as interacting with the detection method or adjusting the GAN directly. In this paper, we introduce a novel class of simple counterattacks that overcomes these limitations. In particular, we show that an adversary can remove indicative artifacts, the *GAN fingerprint*, directly from the frequency spectrum of a generated image. We explore different realizations of this removal, ranging from filtering high frequencies to more nuanced frequency-peak cleansing. We evaluate the performance of our attack with different detection methods, GAN architectures, and datasets. Our results show that an adversary can often remove GAN fingerprints and thus evade the detection of generated images.**

I. Introduction

Generative adversarial networks (GANs) are powerful learning models for synthesizing digital media [10]. They enable generating images and videos that look astonishingly real. For example, the model StyleGAN can generate portrait photos that are not recognizable as synthetic to the human eye [17]. Although GANs have legitimate applications, such as content generation for games and videos [e.g., 18, 25, 32], their ability to create forged images—so called *deep fakes*—resembles a prime tool for misuse, for example, as part of propaganda and disinformation campaigns [5, 27, 30].

Prior work has successfully established different methods for detecting deep-fake images using unique artifacts that GANs leave in the data [e.g., 9, 14, 21, 33, 36, 37]. In particular, the frequency domain of images has proven to be useful for this task, allowing an almost perfect detection [9]. As a result of this performance, different counterattacks have been developed that allow evading the detection of generated images [4, 6, 13]. However, from the adversary's perspective, these attacks still require certain conditions to hold, such as interaction with the detection method or direct adaptation of the GAN model, which limits their practicality.

In this paper, we introduce a novel class of simple counterattacks that overcomes these limitations. These attacks build on the concept of a *GAN fingerprint*, a consistent frequency pattern that characterizes the generation process similar to a camera fingerprint in digital forensics. By identifying and removing this fingerprint from generated images, our attack obstructs frequency-based detection approaches. The fingerprint removal requires no adaption of the GAN model and is agnostic to the detection method. Figure 1 illustrates this concept: The adversary first generates multiple images, estimates the resulting GAN fingerprint (upper row), and finally removes it from a target image (lower row).

The removal of a GAN fingerprint, however, is not a trivial task, as generation artifacts manifest in different frequency bands and patterns. As a consequence, we develop four variants of our attack, gradually increasing their sophistication. We start by simply removing high frequencies from images. This variant is surprisingly effective if the GAN fingerprint is located in high-frequency bands, yet it also affects image details. As a remedy, the second variant targets the fingerprint more precisely by removing the differences between the mean frequency spectra of fake and natural images. The third variant refines this approach and only removes peaks from the frequency differences. Finally, the last variant uses a regression model to estimate discriminative patterns in the frequency spectra.

We empirically evaluate the performance of these four attack variants with different detection methods, GAN architectures, and datasets. In particular, we employ the detection method by Joslin and Hao [14] and two learning-based classifiers by Frank et al. [9]. Our evaluation shows that the removal of GAN fingerprints misleads all detection methods. While the mean-spectrum attack is highly effective against Joslin and Hao, the removal of high frequencies or frequency peaks evades Frank et al. in most cases. Contrary to our expectations, these simple attack variants are more effective than our learning-based regression attack. All in all, our findings demonstrate that adversaries can evade detection methods with relatively simple means and there is a need for more robust concepts.

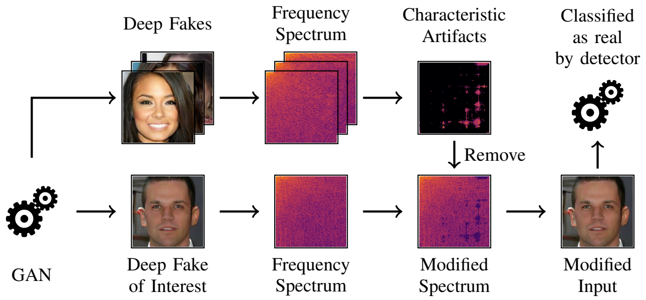


Fig. 1: Illustration of our counterattacks. The adversary calculates the characteristic GAN artifacts in the frequency spectrum and removes this fingerprint to avoid detection.

Contributions. In summary, our contributions are as follows:

- *GAN fingerprints against deep-fake detection.* We show that removing the characteristic artifacts of GAN images in the frequency spectrum is a simple yet effective counterattack against deep-fake detection methods.
- *Manipulation strategies.* We present four methods for modifying the frequency spectrum. They range from removing high frequencies to more nuanced artifact removals.
- *Comprehensive evaluation.* We empirically evaluate our attacks on three detection methods, four GAN architectures, and two datasets. The detection rate from each GAN can be considerably reduced by one of our attacks.

We make the source code and dataset information available under: https://github.com/vwesselkamp/deepfake-fingerprint-attacks.

II. Deep Fake Detection

Approaches for detecting deep-fake images can be broadly divided into two groups: The first group checks the consistency of an image. For instance, inconsistent physical traits can be leveraged [31], such as the pose of the head or facial symmetry of eyes and earrings. Likewise, the color saturation or other disparities in the color components of images can also uncover a deep fake [31]. The second group relies on (invisible) image artifacts that the generation process introduces [9, 14]. Their advantage is that artifacts can be automatically derived for each GAN. This allows for a rather generic identification. Recent work also suggests that artifacts may even transfer between different GANs [33]. In this paper, we focus on these artifact-based approaches.

A. GAN Artifacts and Fingerprints

To provide a first intuition, Figure 2 shows the averaged discrete cosine transform (DCT) spectrum from natural and GAN-generated images, respectively. Two aspects are noticeable: (a) GAN images lead to visible, characteristic artifacts in the frequency spectrum, and (b) these artifacts vary between the different GAN models. For instance, SNGAN induces a grid-like pattern while ProGAN leads to higher values across all frequencies. This simple example underlines that there are clear patterns that differentiate real from GAN images.

The existence of GAN-specific artifacts has been attributed to the up-sampling operations when increasing image resolution [9, 26, 37]. Initially, GANs for image generation [2, 3, 22] used transposed convolution in their up-sampling, which leads to checkerboard artifacts in the spatial domain of images. This occurs when the kernel size is not divisible by the stride by which the kernel moves over the pixels of the low-resolution image. The artifacts created in one layer thus accumulate over several layers and result in patterns in the final image [26]. Hence, recently proposed GANs, such as ProGAN [16], switched to interpolation followed by convolution. While using an interpolation during up-sampling does not produce strong artifacts in the spatial domain anymore, Frank et al. show that different kinds of interpolation still lead to detectable patterns in the frequency domain [9].

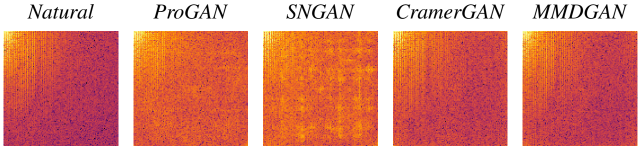


Fig. 2: Mean DCT spectra from real CelebA images and from four GANs on the CelebA dataset. We average the DCT spectrum of 5000 images, log-scale the mean, and cut it to [-10,10], respectively.

These frequency artifacts can be denoted as a *GAN fingerprint* [14, 21], as they are consistently present in images from the same GAN model, but differ between images from different GAN models, similar to a camera fingerprint in digital forensics. This view motivates our counterattacks in Section III that aim at removing or suppressing a GAN fingerprint to bypass a deep-fake detection.

B. Detection Methods

Artifact-based approaches can be further divided into two subgroups: they operate either in the spatial domain [21, 36] or in the frequency domain [8, 9, 12, 14, 28, 37]. A recent comparison by Frank et al. [9] demonstrates multiple advantages of frequency-based approaches, such as a higher accuracy and robustness against image perturbations. For our evaluation, we thus focus on frequency-based approaches and implement the following two detection methods.

First, we consider the fingerprint method by Joslin and Hao [14]. It basically computes a fingerprint by averaging the FFT frequency spectrum of a set of GAN images. The detection is based on computing the cosine similarity between the fingerprint and the FFT spectrum of the image under investigation. Second, we examine the learning-based method by Frank et al. [9]. We consider two models: a Ridge regression and a CNN. Both are trained on the DCT frequency spectrum from natural and generated images. The CNN differentiates five classes (natural images and 4 GAN models), while the regression is a binary classifier that is trained for each GAN individually (see §IV). We choose the regression, since the weights of a regression model have been demonstrated to correspond to periodic patterns in the frequency spectrum. This motivates our fingerprint-based counterattacks that suppress these frequency patterns. The CNN classifier provides the highest detection rate in prior work and thus allows us to test our counterattacks against the current state of the art [9].

III. Counterattacks

We proceed to introduce our novel class of counterattacks. These attacks build on the concept of GAN fingerprints: If a characteristic pattern is present in all generated images, an attacker can try to remove this pattern to evade detection. Such an attack is rather simple to realize. The adversary only has to modify the generated image—adjusting the GAN model is not necessary. Also, the adversary neither requires detailed knowledge of the detection method nor needs to interact with it. As a result, our counterattacks are easy to employ in practice using existing GAN models for generation.

There is, however, a crux: Our evaluation shows that there is *no universal* fingerprint for a GAN that can be simply removed to fool all detection approaches. Instead, each detection method makes use of a different subset of artifacts that affect the detection of fingerprints. Therefore, we derive four different variants of our counterattack with increasing complexity. We start by disturbing the fingerprint through the removal of high frequencies (§III-A) and continue to gradually focus this removal on specific frequency patterns (§III-B).

Notation. Matrices and vectors are written in boldface font. If not stated otherwise, operations on matrices are point-wise. We denote the DCT transformation of a spatial signal $\boldsymbol{X}$ by $\mathcal{D}(\boldsymbol{X}) = \boldsymbol{Y}$, the inverse DCT by $\mathcal{D}^{-1}(\boldsymbol{Y}) = \boldsymbol{X}$. Furthermore, $\boldsymbol{G}$ denotes GAN-generated images, $\boldsymbol{R}$ real images, $\boldsymbol{F}$ fingerprints, and $\tilde{\boldsymbol{G}}$ manipulated GAN images.

Threat Model. We assume a black-box scenario for a counterattack. The adversary has access to a GAN model and uses it to generate deep-fake images. A defender aims at identifying these images using a detection method. The adversary has no inner knowledge of this detection method and cannot interact with the method. Finally, we assume that adapting the GAN model is costly for the adversary. As a result, she focuses on attacks that manipulate the generated images only.

A. Untargeted Fingerprint Removal

Motivated by prior work that establishes the importance of high frequencies for the detection of deep fakes [8, 14, 26, 36], our first attack variant simply removes the high frequency spectrum. In particular, we apply an ideal low-pass filter and set bars of width s of DCT coefficients along the lower and right edges of the spectrum to zero. Figure 3 exemplifies this attack, which we refer to as *frequency-bars attack*.

The attack filters high frequencies from the images. These correspond to details that are less visible for humans and are typically first removed by image compression methods, such as JPEG compression. The intended effect of our attack is similar to *blurring*, which is a typical baseline attack for evading detection in the literature [9, 14, 36]. Yet, the size of the bars s in our attack allows a finer control over the removed information as we demonstrate in §IV.

Although this attack is straightforward to realize and does not require the fingerprint itself, it induces some drawbacks. The attack affects both the fingerprint and the actual image. Moreover, it does not entirely clean a deep fake from artifacts if parts of the fingerprint are located in lower frequencies. As a remedy, we develop more *target-oriented attacks* in the next section that aim at the actual fingerprint.

B. Targeted Fingerprint Removal

We present three attack variants that extract the frequency fingerprint for a GAN model and then suppress it in generated images of this GAN. Figure 3 exemplifies the fingerprints of the presented attacks for the CelebA SNGAN model.

Mean-Spectrum Attack. For this attack, we calculate the difference in the respective mean spectra of natural images and GAN-generated images to determine a fingerprint.

$$\boldsymbol{F}_m = \frac{1}{n}\sum_{i=0}^{n}\mathcal{D}(\boldsymbol{G}_i) - \frac{1}{n}\sum_{i=0}^{n}\mathcal{D}(\boldsymbol{R}_i) \tag{1}$$

As counterattack, we simply subtract the mean fingerprint $\boldsymbol{F}_m$ from a GAN-generated image with strength s:

$$\tilde{\boldsymbol{G}}_i = \mathcal{D}^{-1}(\mathcal{D}(\boldsymbol{G}_i) - s \cdot \boldsymbol{F}_m) \tag{2}$$

Frequency-Peaks Attack. Prior work shows that GAN artifacts are often visible in the frequency domain of images as periodic peaks [9]. We attempt to target these peaks directly by only manipulating the frequency coefficients above a certain threshold. To this end, we again compute the mean spectrum, but now on log-scaled values. As the DCT of an image leads to larger coefficients for low frequencies, log-scaling reduces the emphasis on the low frequencies. We finally execute our manipulations on the non-log-scaled DCT-spectra of GAN-generated images, so that we need to exponentiate the difference. Our *peak* fingerprint $\boldsymbol{F}_p$ becomes:

$$\boldsymbol{F}_p = \exp\left(\frac{1}{n}\sum_{i=0}^{n}\log(\mathcal{D}(\boldsymbol{G}_i)) - \frac{1}{n}\sum_{i=0}^{n}\log(\mathcal{D}(\boldsymbol{R}_i))\right) \tag{3}$$

In this way, frequency patterns become more pronounced in the fingerprint (see Figure 3). We target only the most dominant parts of the pattern: We scale $\boldsymbol{F}_p$ to $[0, 1]$, apply binary thresholding which keeps values larger than a threshold t and sets smaller values to 0, then intensify the kept values with a strength parameter s, and finally clip values to $[0, 1]$ again. The latter avoids switching signs during fingerprint removal. The attack is then given as:

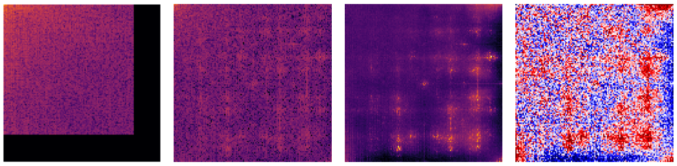
$$\begin{aligned} \tilde{\boldsymbol{G}}_i &= \mathcal{D}^{-1}(\mathcal{D}(\boldsymbol{G}_i)(1 - \tilde{\boldsymbol{F}}_p)) \\ \text{with } \tilde{\boldsymbol{F}}_p &= \text{clip}(s \cdot \text{threshold}(\text{scale}(\boldsymbol{F}_p), t)) \end{aligned} \tag{4}$$

Note that all operations are element-wise. Different from Equation 2, the multiplication reduces the coefficients of the DCT spectrum proportionally to the strength of the fingerprint.

Regression-Weights Attack. For the fourth attack variant, we estimate a fingerprint from weights learned by a regression model. We choose a Lasso regression here, since it pushes the weights of features with little influence on the output towards zero, thus effectively extracting the most relevant features for classification. Moreover, the weights have a direct correspondence to the frequency coefficients, so that a counterattack can directly change the coefficients anti-proportionally to the respective weights. If $\boldsymbol{F}_r$ denotes the regression weights, the counterattack is defined as:

$$\tilde{\boldsymbol{G}}_i = \mathcal{D}^{-1}(\mathcal{D}(\boldsymbol{G}_i)(1 - \tilde{\boldsymbol{F}}_r)) \tag{5}$$

where $\tilde{\boldsymbol{F}}_r$ is clipped, that is, $\tilde{\boldsymbol{F}}_r = \text{clip}(s * \boldsymbol{F}_r)$ with clip reducing the range to $[-1, 1]$.



(a) Frequency bars (b) Mean spectrum (c) Peaks (d) Regression

Fig. 3: Counterattacks for CelebA SNGAN. Plot (a) shows the removal of high-frequency bands; Plot (b)–(d) show the fingerprints that are suppressed. Note that plot (c) shows the fingerprint before applying the threshold.

IV. Evaluation

We proceed to empirically evaluate our counterattacks against deep-fake detection methods. First, we show that the detection rate of deep fakes from each GAN model can be considerably reduced by one of the attack variants (§IV-B) while having only a minor visible impact on the image (§IV-C). Second, we demonstrate that our counterattacks achieve a higher success rate than previously used perturbation-based attacks (§IV-D).

A. Experimental Settings

Dataset and GAN Models. We adopt the experimental setup from prior work [9, 36]: we evaluate four GAN architectures (ProGAN, SNGAN, MMDGAN, CramerGAN) where each is trained on two datasets of natural images (CelebA [20] and LSUN bedrooms [35]), respectively. In total, this setup leads to 8 different combinations of architecture and dataset. The images have a size of $128 \times 128 \times 3$ pixels. Further information about the dataset can be found in our github paper repository.

Deepfake Detectors. As described in §II, we consider multiple detection methods. Table I summarizes the setup. Note that we obtain one detector for each dataset in the multi-class setting, while a binary classifier requires the creation of a single detector for each combination of architecture and dataset.

Detection	Type	Domain
Joslin and Hao [14]	Binary	Frequency (Fourier)
Frank et al. [9] CNN	Multi-class	Frequency (DCT)
Frank et al. [9] Regression	Binary	Frequency (DCT)

TABLE I: Detection setup. Multi-class has five classes {ProGAN, SNGAN, CramerGAN, MMDGAN, Natural}.

To assess the efficacy of our counterattacks, we first compute the accuracy of the detection methods for unmodified deep-fake images. Table II presents the accuracy for each setup. While the approach by Frank et al. [9] exhibits an almost perfect detection rate, the performance of Joslin and Hao [14] varies significantly for different GAN architectures, yielding the best detection rate for SNGAN.

Calibrating Fingerprints. We extract the fingerprints for our attacks on a separate hold-out dataset. The threshold t for the frequency-peaks attack is determined for each GAN model on this set through a simple grid search. For the regression-weights attack, we retrieve the weights for the fingerprint by training a Lasso regression on the hold-out dataset.

Evaluation Measures. We evaluate the performance of counterattacks in terms of *attack success rate* and *image quality*. In particular, we measure the attack performance as the fraction of generated images classified as natural. Note that we aim at a targeted attack in the multi-class setting: an attack is only counted as successful if the detection method misclassifies an image as natural rather than just assigning the wrong GAN class. Furthermore, we measure the visual quality in terms of the Peak Signal to Noise Ratio (PSNR), which is a commonly used metric in image processing [29]. After visual inspection, we consider a PSNR value of 30dB as an acceptable lower bound for the image quality.

B. Attack Success Rate

In the first experiment, we investigate whether the presented counterattacks allow modifying a deep fake such that it is misclassified as a natural image. To this end, we apply the counterattacks on 1,000 images from each GAN model against the three detection methods. Each attack is calibrated using the strength s so that the average PSNR of the 1,000 manipulated images is 30dB.

Results. Table II shows the performance of all attacks with an image quality fixed at 30dB. The attacks reduce the detection rate considerably, demonstrating that deep fakes can be manipulated with fingerprint information only, so that they are classified as actual images.

Attack Analysis. To gain more insights into these results, we first examine the *frequency-bars attack* and its effectiveness against the considered detection methods. Despite its simplicity, the attack is highly successful against the CNN-based classifier and the regression model by Frank et al. These results suggest that the two classifiers mainly rely on information stored in the high-frequency bands for their decisions. In contrast, the attack only provides low success rates against the detector of Joslin and Hao, indicating that low-frequency artifacts are also relevant in the approach.

Interestingly, we obtain the exact opposite results for the *mean-spectrum attack*. This attack works considerably well against Joslin and Hao and can precisely remove the detected pattern. However, it fails to circumvent the classifiers by Frank et al. We attribute the low success rate to the fact that the classifiers operate on log-scaled spectra, while the attack only performs non-scaled manipulations, thus ignoring the peculiarities of the classifiers.

This intuition is further strengthened by the results obtained for the *peak-extraction attack*, which relies on the log-scaled spectra to calculate the fingerprints. The success of the peak-extraction attack, however, depends on the setup: it works almost perfectly against ProGAN and SNGAN, which show strong peaks throughout the spectrum. To confirm that the extracted peaks are accurate for each GAN instance, we perform an additional experiment, in which we cross-remove the fingerprint of individual GAN-instances from images of other GANs. Indeed, we find that removing their own fingerprint results in a more successful attack for each classifier.

				Our counterattacks (success rate)				Baseline perturbations (success rate)			
Dataset	Detection	GAN Model	Accuracy	Frequency bars	Mean spectrum	Peak Extraction	Regression	Cropping	Noise	Blurring	JPEG
LSUN	Joslin	ProGAN	56.7%	69.20%	**96.4%**	71.60%	65.80%	75.70%	74.20%	76.40%	75.50%
		SNGAN	97.8%	13.40%	73.5%	4.70%	4.40%	**95.60%**	10.70%	20.30%	17.10%
		CramerGAN	55.5%	50.80%	**91.6%**	48.10%	47.80%	55.20%	52.90%	56.20%	56.80%
		MMDGAN	57.4%	47.00%	**82.6%**	42.90%	41.90%	54.00%	47.00%	49.70%	50.50%
	CNN	ProGAN	99.0%	89.6%	0%	**92%**	0.1%	12.7%	0%	54.2%	25.2%
		SNGAN		**91.8%**	0%	1.4%	0%	7.3%	0%	56.7%	10.1%
		CramerGAN		**91.1%**	0%	0%	0%	0.3%	0%	62.9%	8.7%
		MMDGAN		**90.8%**	0%	0%	0%	0.2%	0%	56.1%	13.2%
	Regression	ProGAN	91.8%	**100%**	10.4%	**100%**	32.9%	5.1%	82.6%	**100%**	61.5%
		SNGAN	98.9%	**100%**	0%	**100%**	1.7%	24.1%	25.8%	95.3%	13.2%
		CramerGAN	99.1%	**100%**	0%	2.9%	7.9%	35.5%	49.5%	99.5%	80.8%
		MMDGAN	99.3%	**100%**	0.4%	1.2%	57.6%	71.5%	47.7%	99.9%	91%
CelebA	Joslin	ProGAN	79.2%	83.40%	**100.00%**	29.50%	28.40%	84.50%	43.80%	72.20%	69.00%
		SNGAN	95.9%	85.20%	**99.60%**	13.20%	6.40%	96.10%	20.40%	68.00%	66.40%
		CramerGAN	61.3%	73.80%	**95.40%**	53.30%	53.00%	80.80%	61.40%	71.70%	69.20%
		MMDGAN	57.8%	70.30%	**92.30%**	69.10%	69.10%	85.80%	76.90%	78.50%	79.30%
	CNN	ProGAN	99.3%	98.2%	0%	**99.9%**	17.9%	8.4%	0%	**100%**	2.7%
		SNGAN		**100%**	0%	**100%**	1.4%	3.8%	0%	**100%**	1.5%
		CramerGAN		93.1%	0%	0.8%	0%	10.5%	0%	**100%**	2.4%
		MMDGAN		99.5%	0%	0%	0%	25.9%	0.1%	**100%**	3.2%
	Regression	ProGAN	93.3%	20.8%	0.2%	**100%**	73.3%	13.8%	76.2%	85.1%	56.7%
		SNGAN	96.7%	64.7%	0%	**100%**	0.7%	0.6%	60.5%	72.9%	22.4%
		CramerGAN	97.4%	**100%**	0.8%	72.1%	99.9%	53.4%	36.7%	84.2%	47.8%
		MMDGAN	97.3%	97.7%	2.2%	99.1%	**99.4%**	39.1%	38.1%	83.4%	87.1%

TABLE II: The accuracy of deep-fake detection and the success rate of our counterattacks & baseline perturbations for evading the detection—per dataset, detection method, and GAN model. The detection accuracy is computed on 1,000 natural & 1,000 generated images with a binary classifier, and 1,000 natural & 4,000 generated images (1,000 of each GAN model) in a multi-class case. In terms of image quality, the attacks are calibrated to a PSNR value of 30 dB.

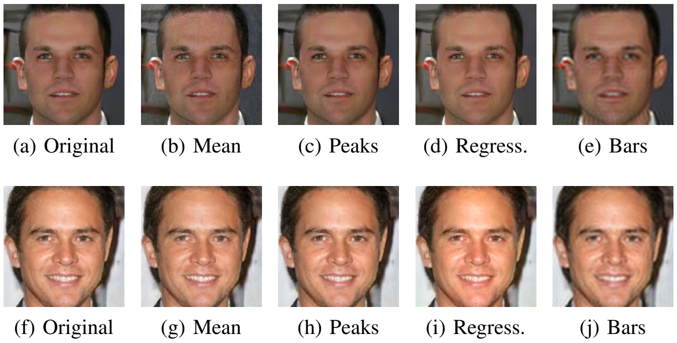
To our own surprise, the *regression-weights attack* is rarely successful—even against a regression model. Our analysis shows that the computed fingerprints exhibit patterns across the entire frequency spectrum, so that the attack also manipulates lower frequency bands. While effective as attacks alone, these manipulations lead to a substantial decrease in image quality and weaken the overall performance.

C. Image Quality

The manipulations performed by our counterattacks in the frequency domain may lead to visible artifacts in the spatial domain. As these artifacts might reveal the attack and provide new ground for detection, we also analyze how much the different counterattacks affect the overall image quality.

Figure 4 shows two representative examples of deep-fake images modified by the different counterattacks at a fixed PSNR of roughly 30dB. While all attacks affect the image quality only slightly, the peak extraction preserves the image details particularly well. Note, however, that this attack yields only moderate success rates (see Table II). The frequency-bars attack, in contrast, introduces more visible artifacts, but the attack also provides good results despite its simplicity, achieving the highest success rates against two of the detection methods. Moreover, we find that its high success rates remain stable even for better PSNR values of up to 37dB, where artifacts are rarely visible anymore.

Overall, these results show that all attack variants are effective with minor impact on the visual quality in most cases. Even for the frequency-bars attack, its impact on the image quality is acceptable on the examined data.


(a) Original (b) Mean (c) Peaks (d) Regress. (e) Bars

(f) Original (g) Mean (h) Peaks (i) Regress. (j) Bars

Fig. 4: Modified deep-fake examples from CelebA SNGAN (a-e) and CelebA ProGAN (f-j). All attacks are performed with a fixed image quality of 30dB.

D. Comparison to Image Perturbations

In the next experiment, we compare our counterattacks with image perturbations that prior work used to test the robustness of detection methods [9, 14, 36]. In particular, we implement cropping, noise addition, blurring, and JPEG compression. We again execute the perturbations with such a strength that the image quality drops to about 30dB on average.

Table II shows attack success rates of the considered image perturbations. While blurring also achieves a high success rate across all settings, the other perturbations show mixed success rates that depend on the respective dataset, detection method, and GAN class/model. In comparison, our counterattacks are more effective, which motivates their usage as additional baselines in future work.

E. Summary of Results

In summary, our experiments demonstrate the effectiveness of the proposed counterattacks. The different variants outperform previous perturbation attacks, without affecting the visual quality significantly at the same time. However, we find that their success depends on various factors, so that a single universal attack strategy does not exist. For instance, while the *mean-spectrum attack* yields the highest success rate against the detection method by Joslin and Hao [14], it is largely ineffective against the other two detectors, where the *frequency-bars attack* and *peak-extraction attack* are most effective here.

V. Discussion

Our evaluation demonstrates how our simple counterattacks impact various deep-fake detectors. Still, there are open questions that we discuss in the following.

A Closer Look on the Frequency Spectrum. Although our presented attacks allow evading the detection in most cases, there is no universal method that is successful in any setup. The success rate can even vary between the different GAN architectures for a respective combination of dataset, detection method, and attack. To better understand the results, we therefore explain the predictions using the example of the CNN-based classifier [9]. We apply LRP, which is a well-established method for analyzing the decisions of various deep neural network architectures [1]. Figure 5 depicts the explanations averaged over 1,000 unmodified deep-fake images.

Our analysis shows that the explanations for the different GAN models and datasets vary—supporting the concept of characteristic GAN artifacts [9]. Amongst others, we find that the relevance of specific frequency bands differs between individual GAN models: while, for instance, the classifier seems to consider the whole spectrum in the case of ProGAN and SNGAN, it mainly focuses on higher frequencies for CramerGAN and MMDGAN. Similarly, the focus on the frequency bands even appears to vary between the datasets for a particular GAN. This finding might also explain the differences we experience between the results on these datasets, as even the same counterattack might yield different success rates depending on the given data (see §IV).

Limitations. We leave counter-defenses to our attacks, the next step in the arms race of attackers and defenders, to future work. For instance, our modified deep-fake images could be added to the training process of deep-fake detectors, similar to adversarial training [11]. Ultimately, iterative research moving back and forth between attacks and defenses likely enables deeper insights on the characteristics of GAN models.

Moreover, we solely focus on frequency-based detectors, which outperform approaches in the spatial domain [9]. However, our preliminary results on attacks against the approach by Yu et al. [36] indicate that frequency-based attacks might be less effective against spatial detection methods. This insight motivates research on fingerprint-based counterattacks in the spatial domain, which we also leave to future research.

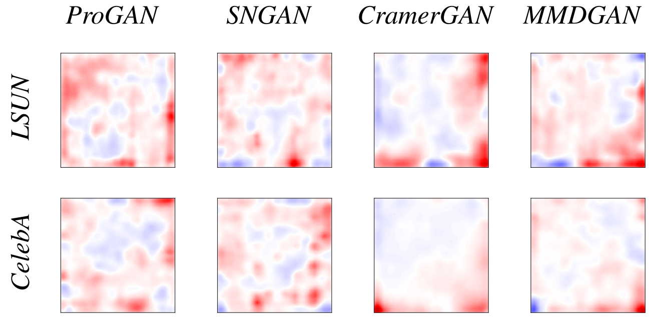


Fig. 5: Mean LRP-explanations in the frequency spectrum for the CNN detector [9]. Red areas correspond to a positive, blue areas to a negative contribution to the deep-fake prediction.

VI. Related Work

The evasion of deep-fake detection methods is an active area of research that can be divided into the following strains: First, an adversary can create an adversarial example of the deep-fake image of interest [4, 19, 23]. Second, the training of the GAN can be directly adapted [7, 15]. For example, Durall et al. [7] show that common upsampling methods prevent models from reproducing the spectral distribution of natural images in the GAN images. Thus, they introduce a spectral regularization term that trains spectrally consistent GANs.

Another line of attacks uses learning-based systems to modify deep fakes [6, 34, 38]. For example, Cozzolino et al. [6] train a GAN to insert the fingerprint of a camera into GAN-generated images while removing the own GAN fingerprint. Neves et al. [24] target high frequencies by using an auto-encoder that encodes an image into a smaller dimensional space before decoding it again, thereby removing unimportant information.

However, Huang et al. [13] state that methods, such as Cozzolino et al. [6] and Neves et al. [24], introduce new artifacts when removing fingerprints. Hence, they propose a shallow reconstruction by learning a dictionary model on natural images, which is a low-dimensional subspace representing these images. A deep-fake is mapped to a representation in the subspace and then reconstructed.

Our approach represents a novel class of attacks. We directly manipulate the frequency spectrum of deep fakes by targeting a GAN fingerprint. The attack operates in a black-box scenario with access to GAN images only. In contrast to prior work, our attacks are conceptually simple and do not require adjusting GANs or training sophisticated learning-based systems.

VII. Conclusion

This paper presents a novel class of simple attacks for bypassing deep-fake detection. The attacks remove GAN artifacts from images directly in the frequency spectrum. Our evaluation shows that depending on the combination of dataset, GAN, and detection method, an adversary can use one of our attacks to mislead the detection. In conclusion, we thus provide evidence that current approaches for detecting deep-fake images are still far from robust and can be evaded easily.

Acknowledgments

We would like to thank Thorsten Holz and Joel Frank for the valuable insights and discussions regarding the presented research idea. The authors gratefully acknowledge funding from the German Federal Ministry of Education and Research (BMBF) under the project BIFOLD (Berlin Institute for the Foundations of Learning and Data, ref. 01IS18025A and ref 01IS18037A) and from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy EXC 2092 CASA-390781972 and the projects 456292433; 456292463.

References

[1] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE*, 2015.

[2] M. G. Bellemare, I. Danihelka, W. Dabney, S. Mohamed, B. Lakshminarayanan, S. Hoyer, and R. Munos. The cramer distance as a solution to biased wasserstein gradients. *arXiv:1705.10743*, 2017.

[3] M. Bińkowski, D. J. Sutherland, M. Arbel, and A. Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations (ICLR)*, 2018.

[4] N. Carlini and H. Farid. Evading deepfake-image detectors with white- and black-box attacks. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2020.

[5] A. Chiu. Facebook wouldn't delete an altered video of Nancy Pelosi. What about one of Mark Zuckerberg? *Washington Post*, 2019.

[6] D. Cozzolino, J. Thies, A. Rössler, M. Nießner, and L. Verdoliva. SpoC: Spoofing camera fingerprints. *arXiv:1911.12069*, 2019.

[7] R. Durall, M. Keuper, and J. Keuper. Watch your up-convolution: CNN based generative deep neural networks are failing to reproduce spectral distributions. *arXiv:2003.01826*, 2020.

[8] R. Durall, M. Keuper, F.-J. Pfreundt, and J. Keuper. Unmasking deepfakes with simple features. *arXiv:1911.00686*, 2020.

[9] J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, and T. Holz. Leveraging frequency analysis for deep fake image recognition. In *Proc. of Int. Conference on Machine Learning (ICML)*, 2020.

[10] 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 Proccessing Systems (NIPS)*, 2014.

[11] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015.

[12] L. Guarnera, O. Giudice, C. Nastasi, and S. Battiato. Preliminary forensics analysis of deepfake images. In *IEEE AEIT International Annual Conference (AEIT)*, 2020.

[13] Y. Huang, F. Juefei-Xu, R. Wang, Q. Guo, L. Ma, X. Xie, J. Li, W. Miao, Y. Liu, and G. Pu. FakePolisher: Making deepfakes more detection-evasive by shallow reconstruction. In *Proc. of the ACM International Conference on Multimedia*, 2020.

[14] M. Joslin and S. Hao. Attributing and detecting fake images generated by known GANs. In *Deep Learning and Security Workshop (DLS)*, 2020.

[15] S. Jung and M. Keuper. Spectral distribution aware image generation. In *Proc. of the AAAI Conference on Artificial Intelligence*, 2021.

[16] 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 (ICLR)*, 2018.

[17] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. Analyzing and improving the image quality of StyleGAN. In *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[18] D. Lee. Deepfake Salvador Dalí takes selfies with museum visitors. https://www.theverge.com/2019/5/10/18540953/salvador-dali-lives-deepfake-museum, 2019.

[19] Q. Liao, Y. Li, X. Wang, B. Kong, B. Zhu, S. Lyu, Y. Yin, Q. Song, and X. Wu. Imperceptible adversarial examples for fake image detection. In *IEEE International Conference on Image Processing (ICIP)*, 2021.

[20] Z. Liu, P. Luo, X. Wang, and X. Tang. Large-scale celebfaces attributes (CelebA) dataset. http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html, 2015.

[21] F. Marra, D. Gragnaniello, L. Verdoliva, and G. Poggi. Do GANs leave artificial fingerprints? In *IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, 2019.

[22] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2018.

[23] P. Neekhara, B. Dolhansky, J. Bitton, and C. C. Ferrer. Adversarial threats to deepfake detection: A practical perspective. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2020.

[24] J. C. Neves, R. Tolosana, R. Vera-Rodriguez, V. Lopes, H. Proença, and J. Fierrez. GANprintR: Improved fakes and evaluation of the state of the art in face manipulation detection. *IEEE Journal of Selected Topics in Signal Processing*, 14(5), 2020.

[25] O. Oakes. 'Deepfake' voice tech used for good in David Beckham malaria campaign. https://www.prweek.com/article/1581457, 2019.

[26] A. Odena, V. Dumoulin, and C. Olah. Deconvolution and checkerboard artifacts. *Distill*, 2016.

[27] D. O'Sullivan. A high school student created a fake 2020 candidate. Twitter verified it. https://www.cnn.com/2020/02/28/tech/fake-twitter-candidate-2020/index.html, 2020.

[28] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. *arXiv:2007.09355*, 2020.

[29] H. T. Sencar and N. Memon, editors. *Digital Image Forensics: There is More to a Picture Than Meets the Eye*. Springer, New York, 2013.

[30] S. Vahia. Deepfake bots create fake nudes of women, aid public shaming and extortion. https://www.moneycontrol.com/news/technology/deepfake-bots-create-fake-nudes-of-women-aid-public-shaming-and-extortion-6081541.html, 2020.

[31] L. Verdoliva. Media forensics and deepfakes: An overview. *IEEE Journal of Selected Topics in Signal Processing*, 14(5), 2020.

[32] J. Vincent. Nvidia has created the first video game demo using AI-generated graphics. https://www.theverge.com/2018/12/3/18121198/ai-generated-video-game-graphics-nvidia-driving-demo-neurips, 2018.

[33] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. 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.

[34] X. Wang, R. Ni, W. Li, and Y. Zhao. Adversarial attack on fake-faces detectors under white and black box scenarios. In *IEEE International Conference on Image Processing (ICIP)*, 2021.

[35] F. Yu, Y. Zhang, S. Song, A. Seff, and J. Xiao. LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv:1506.03365*, 2015.

[36] N. Yu, L. S. Davis, and M. Fritz. Attributing fake images to GANs: Learning and analyzing GAN fingerprints. In *Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[37] X. Zhang, S. Karaman, and S.-F. Chang. Detecting and simulating artifacts in GAN fake images. In *IEEE International Workshop on Information Forensics and Security (WIFS)*, 2019.

[38] X. Zhao and M. C. Stamm. Making GAN-generated images difficult to spot: A new attack against synthetic image detectors. *arXiv:2104.12069*, 2021.