Think Twice Before Detecting GAN-generated Fake Images from their Spectral Domain Imprints

Chengdong Dong [1] Ajay Kumar [1] Eryun Liu [2]
[1] Department of Computing, The Hong Kong Polytechnic University, Hong Kong
[2] College of Information Science and Electronic Engineering, Zhejiang University
chengdong.dong@connect.polyu.hk, ajay.kumar@polyu.edu.hk, eryunliu@zju.edu.cn

Abstract

Accurate detection of the fake but photorealistic images is one of the most challenging tasks to address social, biometrics security and privacy related concerns in our community. Earlier research has underlined the existence of spectral domain artifacts in fake images generated by powerful generative adversarial network (GAN) based methods. Therefore, a number of highly accurate frequency domain methods to detect such GAN generated images have been proposed in the literature. Our study in this paper introduces a pipeline to mitigate the spectral artifacts. We show from our experiments that the artifacts in frequency spectrum of such fake images can be mitigated by proposed methods, which leads to the sharp decrease of performance of spectrum-based detectors. This paper also presents experimental results using a large database of images that are synthesized using BigGAN, CRN, CycleGAN, IMLE, ProGAN, StarGAN, StyleGAN and StyleGAN2 (including synthesized high resolution fingerprint images) to illustrate effectiveness of the proposed methods. Furthermore, we select a spatial-domain based fake image detector and observe a notable decrease in the detection performance when proposed method is incorporated. In summary, our insightful analysis and pipeline presented in this paper cautions the forensic community on the reliability of GAN-generated fake image detectors that are based on the analysis of frequency artifacts as these artifacts can be easily mitigated.

1. Introduction

GAN-based methods can achieve state-of-the-art performance for several computer vision related tasks. They have shown great ability to generate images which do not exist in the real world [8, 32, 38], transfer the style of images [14, 25, 42] and translate text to image [16, 39]. Considering the latent risk associated with the misuse of these fake but real-looklike images, several methods have been proposed to detect such GAN-generated images. Spatial domain methods [36, 40, 41, 44] that directly train large neural network-based detectors have shown to perform well. More recently such fake image detectors based on the artifacts in frequency spectrum of GAN-generated images have been proposed. These detectors require less parameters as compared with the spatial-domain based detectors, and have shown better performance.

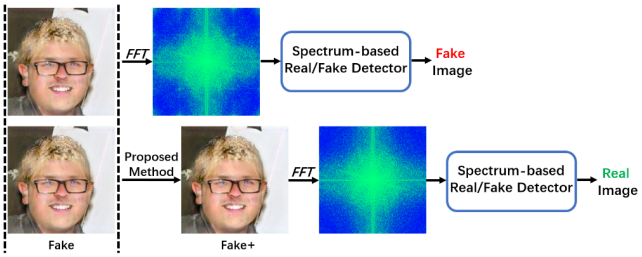


Figure 1. Detectors based on the artifacts in frequency spectrum of GAN-generated images show good performance in recent works. However, these detectors can be compromised when the GAN-generated images are further subjected to our proposed methods.

The main reason for the success of these methods is that the anomalies in the frequency domain representation of GAN-generated images are more pronounced and therefore easy to detect. These anomalies in the spectrum of GAN-generated images can be categorized into two types: abnormal spectral patterns and discrepancy in their power distribution. Some abnormal patterns such as dots and lines are more frequent in the spectra of images generated by CycleGAN [47], StarGAN [13], and StyleGAN [27]. In frequency spectra of BigGAN [10] generated images, cloud-like blurry regions in high-frequency part of spectra are more likely to be observed. In the spectra of synthetic images generated by CRN [12], IMLE [31], ProGAN [26] and StyleGAN2 [28], the artifact patterns have been observed in distinguishing latent shapes. Zhang *et al*. [45] use spec-

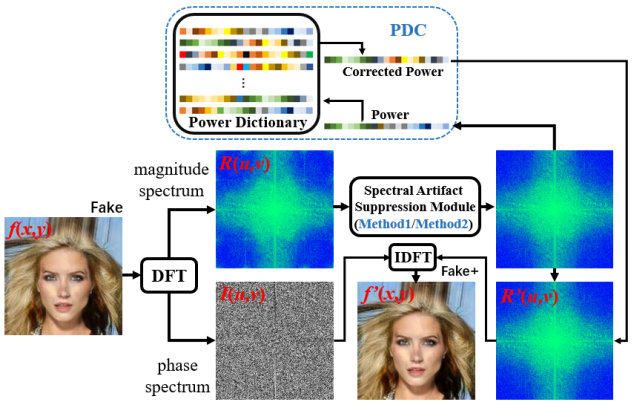


Figure 2. Our pipeline to enhance GAN-generated image to evade popular fake image detectors.

trum as the input to train a classification network and provide comparative performance using a network trained on spatial domain features. Frank *et al*. [23] design a shallow CNN which use DCT spectrum as the input to detect GAN-generated images. It is an exciting approach as it requires less parameters than those in previous works while achieving higher accuracy. Discrepancy in the distribution of spectral power is another type of artifacts. The high frequency part in such power distribution of GAN-generated images can provide important cues of their origins as compared to the real-world images. Therefore, promising attempts have been reported in [19, 20] which train detectors to accurately classify the GAN-generated images, using the cues from the power distribution.

There are some interesting references that analyze the origins of such spectral artifacts and they aim to resolve them. References [19, 20, 23, 45] point out that it largely due to the violation of Sampling Theorem that causes anomaly in the spectra of GAN-generated images. Durall *et al*. [19] introduce spectral loss to correct the power distribution when generator is trained. Chandrasegaran *et al*. [11] analyze the contributions to power distribution from different up-sampling attempts in the last layer of generator. Frank *et al*. [23] analyze the different impacts of three types of up-sampling modules to the anomaly in the patterns of spectrum. Although these works analyze the reason why artifacts exist in frequency spectra of GAN-generated images, the artifacts in spectra are still difficult to alleviate. Therefore, such fake image detectors based on the analysis of frequency domain imprints are widely regarded as robust. On the contrary, our work detailed in this paper develops counterexamples to prove that detectors based on the frequency spectrum may not be robust. The term spectrum in rest of this paper refers to the magnitude spectrum, the term power distribution in this paper essentially refers to spectral power distribution which is computed from magnitude spectrum using Algorithm 1. The spectrum-based fake image detectors, which can also be regarded as the black box models, are still easy to attack. Fig. 1 shows an illustration of the proposed method to generate more effective fake image (Fake+) that can compromise the spectrum-based fake image detectors.

1.1. Contributions

The contributions of our work can be summarized as follows:

- We investigate the mainstream fake image detectors and observe that these spectrum-based detectors mainly rely on the abnormal patterns and power discrepancy of artifacts in their spectra to detect the fake images. Therefore two methods are proposed to mitigate the intensity of artifact patterns in spectra of GAN-generated images, and one method is proposed to correct power discrepancy.

- By incorporating artifact-mitigating strategy followed by the power discrepancy correction, the performance from spectrum-based detectors significantly reduces. It can prove that the spectrum-based detectors are not robust to handle malicious operations on spectra of the fake images. This finding is also significant for the image forensic community. It cautions for their reliance on the spectrum-based fake image detection.

- We also present experimental results to evaluate the accuracy of a spatial domain based detector on the GAN-generated images that are enhanced by proposed methods. Despite performance degradation, such spatial domain based detector is still robust to detect fake images.

1.2. Related Work

Accurate detection of GAN-generated fake images has attracted significant research efforts and promising results have been reported in the literature. These techniques can be categorized into the spatial domain and frequency domain methods, and briefly summarized in the following.

Spatial Domain Methods: Similar to the existence of digital imprints in the images acquired by the real-world cameras, GAN-generated high quality images are expected to present spatial domain imprints that may be visually imperceptible. Marra *et al*. [34] propose a steganalysis method based on photo response non-uniformity (PRNU) patterns. McCloskey *et al*. [35] point out that saturation cues can be utilized to distinguish GAN-generated imagery from camera imagery. Marra *et al*. [33] propose detector based on neural networks to classify the real and fake images and conclude that neural-network-based detectors can perform better than conventional methods for detecting compressed

images. Earlier study from Cozzolino *et al*. [15] notes that the forensic classifiers perform poorly in test dataset where the synthetic images are generated from GANs that are different from those used for training dataset. Therefore they design a few-shot detector for forensic purpose to address this problem. Wang *et al*. [41] train a detection model with only using ProGAN-generated training images to prove that neural-network-based detectors can offer strong generalization capability in distinguishing GAN-generated images.

Frequency Domain Methods: Zhang *et al*. [45] propose a spectrum detector with good generalizing ability which is trained on spectra of images synthesized by a GAN simulator that shares the architectures which are common in most GANs. Frank *et al*. [23] demonstrate that the spectral artifact patterns in fake images can be attributed to the up-sampling operations which are common in all of the popular GAN models. This reference also presents a detailed analysis on three different types of up-sampling modules *i.e*. nearest neighbor up-sampling, bilinear up-sampling, binominal up-sampling and their influence on resulting spectrum. Besides such direct use of the spectrum for training classifier models, several references propose methods that can detect fake images from the analysis of their spectral power distribution. Dzanic *et al*. [20] claim that there is systematic shortcoming of GANs in replicating the attributes of high-frequency modes, and they propose a detector whose cross-model accuracy can reach up to 99.2%. Durall *et al*. [19] suggest that GAN-generated images always possess inevitably higher power distribution in the high frequency portion because of the intrinsic characteristics of convolutional neural networks. They propose simple detectors based on SVM and k-means power distribution classifier and achieve high detection accuracy. Three convolutional layers consisting of 5×5 kernels are also introduced to suppress the power distribution discrepancy in the high-frequency region of the spectrum. However, this method requires carefully selected quantity and size of convolutional layers at the last layer of generator, and is not always very successful in suppressing the higher frequency components. On the other hand, Chandrasegaran *et al*. [11] have recently argued that such abnormal power distribution can actually be avoided by adjusting the up-sampling operation for the last layer of generator. They list several counterexamples to underline the limitation for detecting GAN-generated images from their power distribution discrepancy. However, their proposed method requires careful and manually adjustment of the type and quantity of the up-sampling modules. Therefore it is still difficult to mitigate spectral anomalies for most of the GANs.

2. Proposed Methods

In this section we provide details of the developed methods to correct spectral domain anomalies in the GAN generated images. We firstly introduce SpectralGAN in Sec. 2.1 which is followed by another approach using spectrum difference normalization in Sec. 2.2. Sec. 2.3 introduces a dictionary-based method to correct the power distribution discrepancy in the generated images. As can be observed from Fig. 2, the GAN-generated image $f(x, y)$ can be decomposed into magnitude spectrum $R(u, v)$ and phase spectrum $I(u, v)$ following the Fourier transform. After post-processing $R(u, v)$ with the spectral artifact pattern suppression module, followed by dictionary-based power correction (PDC), artifact-free spectrum $R'(u, v)$ is generated from $R(u, v)$. Finally we use inverse Fourier Transform to recover the $f'(x, y)$ to evade the popular detectors.

2.1. SpectralGAN

The SpectralGAN is the abbreviation of the proposed GAN-based module to mitigate the spectral artifacts in the input images and is also referred to as Method 1 in Sec. 3.

If we regard the domain of spectra of GAN-generated images which contain artifacts as domain A and the spectra of real-world images without such artifacts as domain B, then the task here is to learn the mapping relationship from domain A to domain B. In such scenario, it is natural for us to consider CycleGAN based network architecture to realize such domain transfer. However, our attempts to incorporate original CycleGAN to alleviate such spectral artifacts are not successful. It can be observed in Fig. 3a which illustrates obvious difference in the power distribution of spectra after such domain transfer attempts using the original CycleGAN, and the spectra of real images. We draw Fig. 3 using the fingerprint class in dataset. Further, we note that it is necessary to stabilize the network to ensure that the maximum value in spectrum unchanged, else the training of SpectralGAN can easily collapse. Therefore, we introduce power loss and max loss detailed in the following part.

Optimization Objectives: Given training samples $\{a_i\}$ and $\{b_i\}$ which belong to domain A and domain B respectively, we define $a \in \{a_i\}$, $b \in \{b_i\}$. We denote the mapping function from domain A to domain B as G, and the mapping back from B to A as F. Function D_A aims to distinguish between the actual spectra from domain A and the correspondingly mapped spectra from domain B, while the function D_B aims to distinguish between the spectra from domain B and the correspondingly mapped spectra from domain A. In addition to the conventional L_{GAN}, L_{cyc} and L_{identity} incorporated in CycleGAN, we introduce power loss L_{power} which aims to help regularize the range of spectral power distribution, and max loss L_{max} which is used for ensuring the constraint that the maximum value of spectra remains unchanged during domain transfer A→B, B→A, A→B→A and B→A→B. Therefore the total loss function $L\left(G, F, D_A, D_B\right)$ is stated in Eq. (1).

$$
\begin{aligned}
L\left(G, F, D_A, D_B\right) = & L_{\text{GAN}}\left(G, D_B, A, B\right) \\
& + L_{\text{GAN}}\left(F, D_A, B, A\right) \\
& + \lambda_1 L_{\text{cyc}}(G, F) + \lambda_2 L_{\text{identity}}(G, F) \\
& + \lambda_3 L_{\text{power}} + \lambda_4 L_{\max}
\end{aligned} \tag{1}
$$

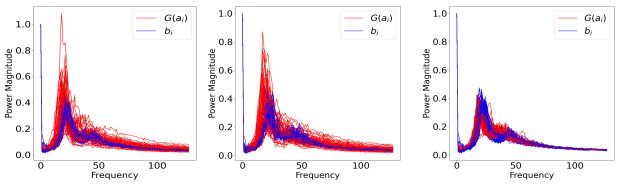
where λ_1, λ_2, λ_3, λ_4 represent the weights for L_{cyc}, L_{identity}, L_{power}, $L_{\max}$ respectively.

We aim to optimize G^*, F^*:

$$
G^*, F^* = \arg\min_{G,F} \max_{D_A, D_B} L\left(G, F, D_A, D_B\right) \tag{2}
$$

The max loss is stated in Eq. (3):

$$
\begin{aligned}
L_{\max} = & \|G(a)[g][h] - a[g][h]\|_1 + \|F(b)[p][q] - b[p][q]\|_1 \\
& + \|F(G(a))[g][h] - a[g][h]\|_1 \\
& + \|G(F(b))[p][q] - b[p][q]\|_1 \\
& (g, h = \arg\max_{g,h} a[g][h], \quad p, q = \arg\max_{p,q} b[p][q])
\end{aligned} \tag{3}
$$



(a) CycleGAN (b) spectral loss (c) L_{power} & $L_{\max}$

Figure 3. The comparative spectral power distributions of different optimization constraints. The blue curves illustrate the power distributions of samples b_i in domain B, the red curves illustrate power distributions of spectra which are mapped from samples a_i in domain A using function G.

Power Loss: Durall *et al.* [19] introduce spectral loss into generator of GAN to constrain the energy distribution of generated images. Although the spectral loss incorporated for spatial domain features has shown to offer great performance to restrict power distribution, when we introduce spectral loss to constrain the power distribution directly on the spectral domain features, it fails to constrain such power distribution well. It can be observed in Fig. 3b where the power distributions of spectra after introduction of such optimization constraint are different from those of real images. Therefore, we propose to incorporate the power loss in Eq. (4) for constraining the spectral power distribution instead of spectral loss:

$$
L_{\text{power}} = \|P(G(a)) - \boldsymbol{l}_{mean}\|_F \tag{4}
$$

where $\boldsymbol{l}_{mean}$ is the average power distribution of real images, $\|\cdot\|_F$ denotes Frobenius norm, computation of spectral power distribution which follows Algorithm 1 is denoted as P. For $M \times M$ images, the size of $\boldsymbol{l}_{mean}$ and $P(G(a))$ is $1 \times \lfloor\sqrt{2}M\rfloor$.

Algorithm 1 Computation of Spectral Power Distribution

Input: Spectrum $\boldsymbol{S}$, whose size is $M \times M$
Output: Spectral Power Distribution Feature $\boldsymbol{l}$, which is initialized as $1 \times M$ zero vector

for i in range(M) **do**
  for j in range(M) **do**
    index $= \lfloor\sqrt{(i-0.5M)^2 + (j-0.5M)^2}\rfloor$
    $\boldsymbol{l}$[index] $= \boldsymbol{l}$[index] $+ S[i][j]$
  end for
end for
$\boldsymbol{l} = \boldsymbol{l}/\boldsymbol{l}[0]$

Algorithm 2 Spectrum Difference Normalization

Input: Spectrum $\boldsymbol{S}$, average spectrum $\overline{\boldsymbol{R}}$ of real-world images in training dataset, average spectrum $\overline{\boldsymbol{G}}$ of GAN-generated images in training dataset
Output: Spectrum $\boldsymbol{S}'$

The spectrum difference is denoted as $\Delta = \overline{\boldsymbol{G}} - \overline{\boldsymbol{R}}$
$\boldsymbol{S}' = \boldsymbol{S} - \Delta$
return $\boldsymbol{S}'$

We can observe from Fig. 3c that the spectral power distributions are successfully restricted into normal distribution when power loss and max loss are all incorporated.

Backbone of Generator and Discriminator: The backbone of generator for SpectralGAN is Nested UNet [46]. In Fig. 4 we present the visualization of intermediate features during forward propagation in this generator. This figure helps us to ascertain that the generator has effectively learnt to suppress the spectral artifacts in such skip-connected architecture, without influencing the salient visual details that generally reside in low frequency part of spectrum. The definition and variables $x_{00}, x_{01}, \ldots, x_{40}$ employed in this visualization are the same as defined in Nested UNet [46]. The size of input and output features is $1 \times 256 \times 256$, the size of $x_{00} \sim x_{04}$ is $64 \times 256 \times 256$, the size of $x_{10} \sim x_{13}$ is $128 \times 128 \times 128$, the size of $x_{20} \sim x_{22}$ is $256 \times 64 \times 64$, the size of x_{30}, x_{31} is $512 \times 32 \times 32$, the size of x_{40} is $1024 \times 16 \times 16$. For visualization, we calculate the average value of the first channel. We adopt the same architecture for the discriminator as in PatchGAN [25], which helps the network converge quickly with relatively small number of parameters and larger receptive fields.

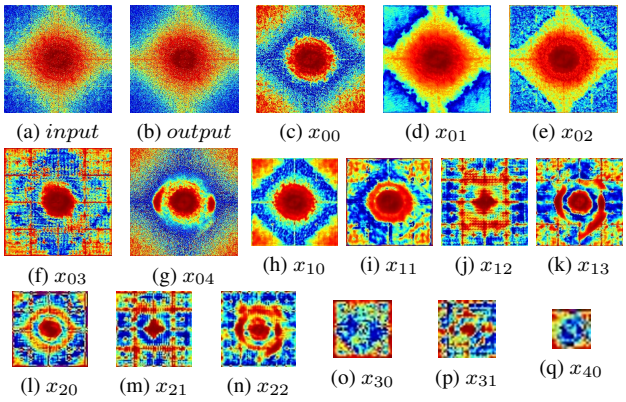


Figure 4. Intermediate features during forward propagation .

2.2. Spectrum Difference Normalization

In this section, we propose to incorporate the differences in the spectrum of two domains to mitigate the artifacts in the spectrum and this approach is referred to as Method 2 in Sec. 3 on experiments. Although it is hard to determine the intensity and exact location of artifacts, simply by subtracting the average of the spectrum differences Δ, *i.e.* between the fake images and real images, on the spectra with artifacts can achieve good results. The details of implementation of this approach is summarized in Algorithm 2.

2.3. Power Distribution Correction

In this section, a dictionary-based power distribution correction (PDC) method is proposed to further correct the power distribution discrepancy. Although we have considered two methods to mitigate the artifacts in spectra, they may fail for some images in ensuring the realness of the power distribution, and our extensive experiments in Tab. 1 reveal that such power distributions can still be detected by a linear classifier. Thus, it is necessary to introduce power distribution correction method to address this problem. We firstly construct a dictionary of power distribution features of real-world images, then we detail the steps to correct the power distribution using this approach.

Power Dictionary Construction: This dictionary consists of a number of 1-dimensional power distribution features in the available real-world images dataset. The power dictionary features from spectra of real-world images are computed using Algorithm 1. The power distribution dictionary $\boldsymbol{D}$ which contains N features can be represented as follows:

$$\boldsymbol{D} = \{D_1, D_2, \ldots, D_N\}, D_i = [D_{i_1}, D_{i_2}, \ldots, D_{i_k}]$$

$$\text{where } i = 1, 2, \ldots, N \quad \text{and} \quad k = \lfloor \sqrt{2}M \rfloor$$

Dictionary-based Correction: Given a spectrum processed by the spectral artifacts suppression module as shown in Fig. 2, its power distribution feature can be corrected from the dictionary $\boldsymbol{D}$ in the following Algorithm 3.

Algorithm 3 Dictionary-based Power Correction

Input: Spectrum $\boldsymbol{S}$, whose size is $M \times M$
Output: The spectrum $\boldsymbol{S}'$ with corrected power distribution

1: Calculate the power distribution P_s of $\boldsymbol{S}$ following Algorithm 1. $P_s = [P_{s_1}, P_{s_2}, \ldots, P_{s_k}]$
2: We select the power distribution feature D_r in the dictionary, $r = \arg\min_i \sum_{j=1}^{\frac{1}{4}k} \left(P_{s_j} - D_{i_j}\right)^2$
3: **for** i in range(M) **do**
4: **for** j in range(M) **do**
5: index $= \lfloor \sqrt{(i-0.5M)^2 + (j-0.5M)^2} \rfloor$
6: $\boldsymbol{S}'[i][j] = S[i][j] \times \frac{D_r[\text{index}]}{P_s[\text{index}]}$
7: **end for**
8: **end for**

3. Experiments and Results

In Sec. 3.1 we briefly detail two recent and popular spectrum-based detectors along with one spatial feature-based detector to evaluate the performance of the methods proposed in our paper. In *supplementary file* we provide implementation details for Method 1, Method 2 and PDC. We analyze the effectiveness of the proposed methods in Sec. 3.2. Finally, several examples from challenging samples are presented in Sec. 3.3 to outline limitation of proposed methods. The spectra used for training and testing Method 1, Method 2 and CNN-based spectrum detectors are all log-scaled and normalized to $[-1, 1]$ range, the power distribution features used for constructing power dictionary and training SVM classifier are computed from image spectra which have not been log-scaled and normalized.

Introduction to Dataset: The dataset contains nine classes, each class includes two sub-classes: GAN-generated images, and real-world images used for training the corresponding type of GAN. Similar to as in reference [41], synthetic images generated using BigGAN, CRN, CycleGAN, IMLE, ProGAN, StarGAN, StyleGAN, StyleGAN2 and the corresponding real images used for training these GANs consist the first 8 class. We additionally select real and GAN-generated fingerprint images as the last class of the dataset to evaluate effectiveness of proposed methods, since the spectra of fingerprint images are significantly different from other real-world images and such images are widely used for a range of security and e-governance applications [17, 18, 29, 37]. Please see *supplementary file* for more details on this database. The fake images in BigGAN $\sim$ StyleGAN2 classes are acquired from [1, 41]. Further we also acquire the fake images of BigGAN, CycleGAN, StarGAN, StyleGAN and StyleGAN2 classes that are provided in [2–6]. For fingerprint class, synthetic fingerprints are generated from skeletons following the method proposed in [43]. We generate fake fingerprint images from skele-

Metric	BigGAN	CRN	CycleGAN	IMLE	ProGAN	StarGAN	StyleGAN	StyleGAN2	Fingerprint
Original fake images	78.81	90.37	90.55	72.40	92.21	88.25	74.23	76.98	100.0
Method 1 + PDC	51.00	51.88	51.90	52.11	51.72	50.60	51.73	51.39	50.19
Method 1	58.34	68.44	62.15	61.32	72.95	68.83	65.48	60.31	55.58
Method 2 + PDC	50.47	51.80	51.68	54.98	50.56	**50.33**	50.95	50.72	50.78
Method 2	57.03	69.42	64.15	62.65	73.70	67.58	63.80	61.44	60.16
PDC	**50.22**	**50.20**	**51.35**	**50.43**	**50.15**	50.63	**50.08**	**50.34**	**50.12**

Table 1. Results from SVM Classifier using the power distribution (accuracy in %)

Metric	BigGAN	CRN	CycleGAN	IMLE	ProGAN	StarGAN	StyleGAN	StyleGAN2	Fingerprint
Original fake images	94.59	96.32	99.48	95.11	97.62	98.08	85.48	92.55	100.0
Method 1 + PDC	68.44	81.09	**53.30**	61.79	64.58	**54.73**	**64.05**	69.28	**51.09**
Method 2 + PDC	**60.34**	**72.67**	59.18	**57.79**	**61.32**	60.28	67.19	**62.63**	54.55
PDC	93.28	91.50	94.33	92.68	91.49	93.85	80.13	87.02	100.0

Table 2. Results from shallow CNN-based spectrum detector (accuracy in %)

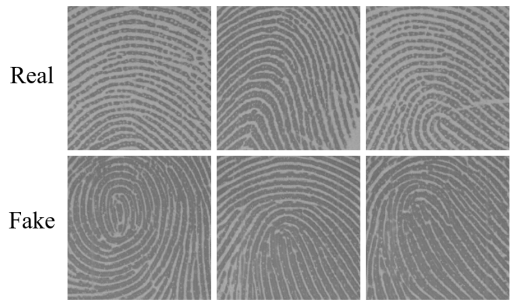


Figure 5. Real and GAN-generated fake fingerprint images.

tons followed by style transfer to synthesize real look-like fingerprint images. Sample real fingerprint images used for training the style transfer models are illustrated in the first row of Fig. 5 while the GAN-generated fake fingerprint image samples are illustrated in the second row of Fig. 5.

All the images are resized to 256×256, 80% of the GAN-generated and real images in each class are used for training and the rest of the images is used for testing.

3.1. Fake Image Detectors

SVM Classifier using the Power Distribution: This spectral power distribution based fake image detector uses a traditional supervised SVM classifier. We recreate this detector by Durall *et al*. [19] and ensure that the proposed methods can effectively correct the power distribution discrepancy and successfully challenge this detector. The power distribution features of size 1×180 are computed using Algorithm 1. We train a SVM classifier using power distribution features from 82210 GAN-generated images and 82210 real images from 8 classes (excluding the fingerprint class) in the training dataset as detailed in the *supplementary file*, the experimental results of BigGAN∼StyleGAN2 in Tab. 1 are tested by this detector. Further, as the power distributions of fingerprint images are different from other real-world images, we independently train another SVM classifier using respective power distribution features from training dataset of fingerprint images. The fake fingerprint images detection results in Tab. 1 are tested by this detector.

CNN-based Spectrum Classifier: Reference [23] presents another promising detector which trains a shallow CNN with only four convolution layers to detect the GAN-generated images using the spectral features. Such shallow CNN trained on entire spectrum has shown to offer significant performance boost as compared to those classifiers trained on raw pixels. Our paper adopts two CNN models to detect the GAN-generated images using the entire spectrum: a shallow CNN which only contains 4 layers and a ResNet18 classifier. The spectra of 88126 GAN-generated images and 89122 real images from nine classes of training dataset are used for training these two detectors.

CNN-based Spatial Domain Classifier: We also use a pretrained model provided in [1, 41] which is trained on ProGAN-generated images and has shown to offer high generalizing capability for detecting fake images generated by unseen GANs. This fake image detector is referred to as detector W, which is able to calculate the realness score of the input RGB image.

3.2. Discussion on Results

Method 1 and Method 2: The experimental results summarized in Tab. 1 illustrate the limitations of Method 1 and Method 2 if these methods are not combined with dictionary-based power correction. Therefore the SVM

Metric	BigGAN	CRN	CycleGAN	IMLE	ProGAN	StarGAN	StyleGAN	StyleGAN2	Fingerprint
Original fake images	95.63	97.06	99.85	96.35	98.18	99.20	89.31	94.91	100.0
Method 1 + PDC	70.88	83.91	**54.20**	63.02	72.63	**53.03**	**68.55**	71.27	**52.09**
Method 2 + PDC	**61.81**	**80.81**	60.53	**57.51**	**66.68**	59.05	69.02	**65.33**	56.51
PDC	93.63	91.35	95.15	91.70	92.91	94.48	78.98	86.97	100.0

Table 3. Results from ResNet-based spectrum detector (accuracy in %)

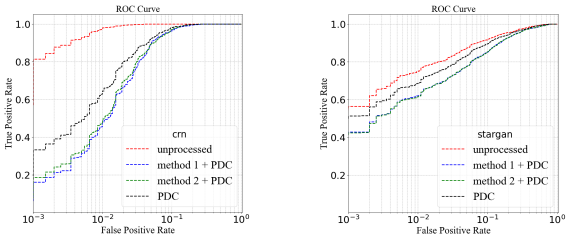
Dictionary with 100k Samples	BigGAN	CRN	CycleGAN	IMLE	ProGAN	StarGAN	StyleGAN	StyleGAN2
Method 1 + PDC	0.9493	0.9245	0.9485	0.9638	0.9409	0.8978	0.9015	0.9391
Method 2 + PDC	0.9510	0.9222	0.9158	0.9627	0.9399	0.9262	0.9239	0.9388
PDC	**0.9611**	**0.9505**	**0.9648**	**0.9654**	**0.9550**	**0.9336**	**0.9566**	**0.9427**
Dictionary with 200k Samples	BigGAN	CRN	CycleGAN	IMLE	ProGAN	StarGAN	StyleGAN	StyleGAN2
Method 1 + PDC	0.9524	0.9289	0.9567	**0.9734**	0.9513	0.9117	0.9131	0.9425
Method 2 + PDC	0.9469	0.9316	0.9390	0.9705	0.9412	0.9285	0.9346	0.9447
PDC	**0.9639**	**0.9522**	**0.9714**	0.9663	**0.9742**	**0.9515**	**0.9580**	**0.9514**
Dictionary with 500k Samples	BigGAN	CRN	CycleGAN	IMLE	ProGAN	StarGAN	StyleGAN	StyleGAN2
Method 1 + PDC	0.9596	0.9302	0.9807	0.9728	0.9685	0.9460	0.9281	0.9554
Method 2 + PDC	0.9518	0.9325	0.9487	0.9713	0.9651	0.9384	0.9383	0.9536
PDC	**0.9721**	**0.9534**	**0.9828**	**0.9756**	**0.9815**	**0.9672**	**0.9617**	**0.9589**

Table 4. Similarity between original fake images and enhanced fake images (fake+) using SSIM scores

classifier is still capable of detecting GAN-generated fake images but with reduced accuracy, which mirrors the justification to introduce the power correction. We report the performance from two CNN-based spectrum detectors in Tabs. 2 and 3. In conjunction with PDC, Method 1 can outperform Method 2 in detecting CycleGAN, StarGAN, StyleGAN and fingerprint classes of fake images, but is outperformed by Method 2 for images in other classes. It is also easy to observe from Tabs. 2 and 3 that both the spectrum-based detectors with two different kinds of CNN architectures, perform quite well in detecting unprocessed GAN-generated fake images and show sharp decline in the accuracy when images are enhanced using the proposed methods. This observation underlines that our proposed methods can be regarded as black box attack to compromise the CNN-based spectrum detectors. Lack of expected features from the spectral artifacts cannot alert the pretrained detector models and thereby disable their capabilities to detect fake images, even without incorporating adversarial attack [9, 21, 22, 24, 30] techniques for compromising such black box model. As can be observed from the results in Fig. 6, by applying the proposed methods the spatial-based fake image detector also illustrates consistently decrease in the detection performance. Therefore it is reasonable to infer that the spatial-domain based detector of [1,41] may also use the frequency domain information while detecting fake images from spatial-domain features.

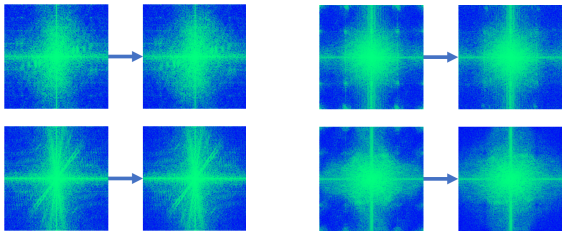
Influence from Power Correction: Our results in Tab. 1 indicate that the dictionary-based power correction is very effective, regardless of whether it is combined with Method 1 or Method 2, to compromise the SVM classifier. The results in the last rows of Tabs. 2 and 3 also indicate that the performances of CNN-based spectrum detector almost remains unchanged if we only incorporate the PDC. The results in these tables reveal that the CNN-based spectrum detectors may not detect the fake images by just using the features from the spectral power distributions. Instead, these detectors more focus on the artifact patterns which mirrors the claim in [23]. The results in Fig. 6 also indicate that the dictionary-based power correction plays a positive role in compromising detection performance of detector W. As compared with the GAN-generated images which are not subjected to the proposed methods, we can note the drop in detection performance when power correction is introduced. As shown from the results in Fig. 6, the performance of spatial-based detector drops further if Method 1 and Method 2 are combined with PDC. However, such phenomenon may not be consistent on other classes of datasets as can be observed from the additional ROC plots which are included in the *supplementary file*.

Influence from the Size of Power Dictionary: Results in Tab. 4 illustrate the structural similarity index measure



(a) CRN (b) StarGAN

Figure 6. ROC performance curves using detector W when proposed methods are evaluated on CRN and StarGAN images.



(a) processed by Method 1 (b) processed by Method 2

Figure 7. Sample results from challenging examples

(SSIM) scores which represent the similarity between the original fake images and the enhanced images (fake+) using three different sizes of such dictionaries. As compared with the other two sets of experiments that combine Method 1 and Method 2 with PDC, only incorporating PDC always achieves the highest SSIM score. As the size of power dictionary increases, the diversity of power distribution features enhances. This can lead to smaller changes in the power distribution of spectra and thereby increasing the SSIM scores.

3.3. Analysis on challenging examples

Challenges for SpectralGAN: We present sample results from challenging examples to ascertain the effectiveness of the proposed methods. We can observe the failure of Method 1 in suppressing the spectral artifact patterns from the examples in Fig. 7a. When the artifact pattern is weak or different from those in the training dataset, Method 1 can decrease its effectiveness in mitigating them. We visualize the samples from spectra of all types of GAN-generated images in training dataset and observe that the artifact patterns of CycleGAN, StarGAN and StyleGAN are visually clear while the artifact patterns from BigGAN, CRN, IMLE and StyleGAN2 images are relatively weaker. This phenomenon may explain the reason for superior performance from Method 1 with PDC, on CycleGAN, StarGAN and StyleGAN, over the Method 2 with PDC as observed from results in Tabs. 2 and 3.

Challenges for Spectrum Normalization: Since the intensity and location of artifact patterns are uncertain, direct subtraction of spectrum differences may lead to two types of errors. The first scenario is that artifacts cannot be sufficiently mitigated. As can be observed from sample results in the first row of Fig. 7b, the artifact pattern can still remain obvious after such normalization. Another failure scenario can arise when the intensity of such artifacts is comparatively weaker and we can observe such example from the sample in the bottom row of Fig. 7b that illustrates intensity at the location of artifacts are smaller than other values in the areas surrounding the artifacts.

4. Conclusions and Further Work

This paper has revisited the effectiveness of the advanced fake image detectors to detect GAN-generated images. We have accordingly developed two methods to mitigate the intensity of patterns from spectral artifacts and one dictionary-based method to correct the spectral power distribution discrepancy in the GAN-generated fake images. Reproducible [7] experimental results presented in Sec. 3 suggest that the proposed methods can effectively mitigate the widely observed spectral artifacts and correct spectral power distribution discrepancy to compromise the spectrum-based detectors for the GAN-generated fake images. The CNN based classifier which uses the entire spectrum for detecting such fake images has still shown its capability in detecting the GAN-generated images, but with significant drop in its performance. The SVM based detector, which locks on power distribution discrepancy, completely fails to detect the GAN-generated fake images that are enhanced by our methods. A spatial-based detector [1, 41] selected in our evaluation however remains robust in distinguishing the enhanced GAN-generated fake images but with degradation in detection performance. We further test the SSIM score between the enhanced images (fake+) and original fake images, which confirms their high similarity. It may be easy to comprehend that why power distribution can be corrected by dictionary-based method in Sec. 2.3. But as for the CNN-based spectrum detectors, one of the reasons that these detectors can be compromised is that they are vulnerable as black-box models to attacker. Since they utilize the artifact information to detect the fake images, we can compromise these models only by removing such artifacts.

In summary, the findings from this research indicate that the detectors based on analysis of fake image spectra may not be robust to detect GAN-generated fake images because the artifacts in such spectra can be easily mitigated by malicious attackers while the respective images can visually look very similar after such enhancement. Therefore significant work is required to develop robust fake image detectors that can reliably detect fake images from sophisticated GAN-based models that pose increasing challenges to our community.

References

[1] https://github.com/peterwang512/CNNDetection. 5, 6, 7, 8

[2] https://github.com/huggingface/pytorch-pretrained-BigGAN. 5

[3] https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix. 5

[4] https://github.com/yunjey/stargan. 5

[5] https://github.com/NVlabs/stylegan. 5

[6] https://github.com/lucidrains/stylegan2-pytorch. 5

[7] Weblink to download the source code for the approach detailed in this paper. https://www.comp.polyu.edu.hk/~csajaykr/deepdeepfake.htm. 8

[8] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017. 1

[9] Bir Bhanu, Ajay Kumar, et al. *Deep learning for biometrics*. Springer, 2017. 7

[10] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 1

[11] Keshigeyan Chandrasegaran, Ngoc-Trung Tran, and Ngai-Man Cheung. 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*, pages 7200–7209, 2021. 2, 3

[12] Qifeng Chen and Vladlen Koltun. Photographic image synthesis with cascaded refinement networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1511–1520, 2017. 1

[13] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul 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. 1

[14] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8188–8197, 2020. 1

[15] Davide Cozzolino, Justus Thies, Andreas Rössler, Christian Riess, Matthias Nießner, and Luisa Verdoliva. Forensictransfer: Weakly-supervised domain adaptation for forgery detection. *arXiv preprint arXiv:1812.02510*, 2018. 3

[16] Bo Dai, Sanja Fidler, Raquel Urtasun, and Dahua Lin. Towards diverse and natural image descriptions via a conditional gan. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2970–2979, 2017. 1

[17] UIDAI dashboard for fingerprint authentication. https://uidai.gov.in/aadhaar_dashboard/auth_trend.php?auth_id=biofingure, Accessed Nov 2021. 5

[18] Office of Biometric Identity Management Department of Homeland Security. https://www.dhs.gov/obim, Accessed Nov 2021. 5

[19] Ricard Durall, Margret Keuper, and Janis Keuper. 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*, pages 7890–7899, 2020. 2, 3, 4, 6

[20] Tarik Dzanic, Karan Shah, and Freddie Witherden. Fourier spectrum discrepancies in deep network generated images. *arXiv preprint arXiv:1911.06465*, 2019. 2, 3

[21] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1625–1634, 2018. 7

[22] Hany Farid. *Fake photos*. MIT Press, 2019. 7

[23] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *International Conference on Machine Learning*, pages 3247–3258. PMLR, 2020. 2, 3, 6, 7

[24] Gaurav Goswami, Akshay Agarwal, Nalini Ratha, Richa Singh, and Mayank Vatsa. Detecting and mitigating adversarial perturbations for robust face recognition. *International Journal of Computer Vision*, 127(6):719–742, 2019. 7

[25] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 1, 4

[26] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 1

[27] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 1

[28] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020. 1

[29] Ajay Kumar. *Contactless 3D Fingerprint Identification*. Springer, 2018. 5

[30] Dongze Li, Wei Wang, Hongxing Fan, and Jing Dong. Exploring adversarial fake images on face manifold. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5789–5798, 2021. 7

[31] Ke Li, Tianhao Zhang, and Jitendra Malik. Diverse image synthesis from semantic layouts via conditional imle. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4220–4229, 2019. 1

[32] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2017. 1

[33] Francesco Marra, Diego Gragnaniello, Davide Cozzolino, and Luisa 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. 2

[34] Francesco Marra, Diego Gragnaniello, Luisa Verdoliva, and Giovanni Poggi. Do gans leave artificial fingerprints? In *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 506–511. IEEE, 2019. 2

[35] Scott McCloskey and Michael Albright. Detecting gan-generated imagery using saturation cues. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 4584–4588. IEEE, 2019. 2

[36] Huaxiao Mo, Bolin Chen, and Weiqi Luo. Fake faces identification via convolutional neural network. In *Proceedings of the 6th ACM Workshop on Information Hiding and Multimedia Security*, pages 43–47, 2018. 1

[37] Federal Bureau of Investigation. Integrated automated fingerprint identification system. https://www.fbi.gov/services/information-management/foipa/privacy-impact-assessments/iafis, Accessed Nov 2021. 5

[38] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 1

[39] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International Conference on Machine Learning*, pages 1060–1069. PMLR, 2016. 1

[40] Shahroz Tariq, Sangyup Lee, Hoyoung Kim, Youjin Shin, and Simon S Woo. Gan is a friend or foe? a framework to detect various fake face images. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, pages 1296–1303, 2019. 1

[41] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8695–8704, 2020. 1, 3, 5, 6, 7, 8

[42] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018. 1

[43] André Brasil Vieira Wyzykowski, Mauricio Pamplona Segundo, and Rubisley de Paula Lemes. Level three synthetic fingerprint generation. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 9250–9257. IEEE, 2021. 5

[44] Ning Yu, Larry S Davis, and Mario Fritz. Attributing fake images to gans: Learning and analyzing gan fingerprints. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7556–7566, 2019. 1

[45] Xu Zhang, Svebor Karaman, and Shih-Fu Chang. Detecting and simulating artifacts in gan fake images. In *2019 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6. IEEE, 2019. 1, 2, 3

[46] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 3–11. Springer, 2018. 4

[47] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei 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. 1