A Closer Look at Fourier Spectrum Discrepancies for CNN-generated Images Detection

Keshigeyan Chandrasegaran    Ngoc-Trung Tran    Ngai-Man Cheung
Singapore University of Technology and Design (SUTD)
{keshigeyan, ngoctrung_tran, ngaiman_cheung} @sutd.edu.sg

Abstract

CNN-based generative modelling has evolved to produce synthetic images indistinguishable from real images in the RGB pixel space. Recent works have observed that CNN-generated images share a systematic shortcoming in replicating high frequency Fourier spectrum decay attributes. Furthermore, these works have successfully exploited this systematic shortcoming to detect CNN-generated images reporting up to 99% accuracy across multiple state-of-the-art GAN models.

In this work, we investigate the validity of assertions claiming that CNN-generated images are unable to achieve high frequency spectral decay consistency. We meticulously construct a counterexample space of high frequency spectral decay consistent CNN-generated images emerging from our handcrafted experiments using DCGAN, LSGAN, WGAN-GP and StarGAN, where we empirically show that this frequency discrepancy can be avoided by a minor architecture change in the last upsampling operation. We subsequently use images from this counterexample space to successfully bypass the recently proposed forensics detector which leverages on high frequency Fourier spectrum decay attributes for CNN-generated image detection.

1. Introduction

With serious concerns over Deepfakes being widely used for malicious purposes [18, 19, 30, 37, 2, 36, 9, 32], detection of deepfake multimedia content has become an important research field. With substantial improvement of CNN-based generative modelling in the recent years [22, 45, 23, 24, 21, 35, 8, 3, 46, 33, 27, 1, 4, 16, 38, 39], it is becoming more and more difficult to assess the “fakeness” of such synthetic content in the RGB pixel space.

1.1. Fourier spectrum discrepancies in CNN-generated images

Recent research suggests that CNN-based generation methods are unable to reproduce high frequency distribution of real images. Existing work tends to conclude that this incompetency is an intrinsic property of CNN-based generative models [12, 10, 25]. While Zhang et al. [44] and Wang et al. [42] report that CNN generated images have frequency artifacts, Dzanic et al. [12] and Durall et al. [10] suggest spectrum discrepancies in high frequency: CNN-generated images at the highest frequencies do not decay as usually observed in real images (Figure 1). In particular,

- Dzanic et al. [12] analyze high frequency of real and deep network generated images, and conclude that “deep network generated images share an observable, systematic shortcoming in replicating the attributes of these high-frequency modes”.

- Durall et al. [10] observe that “CNN based generative deep neural networks are failing to reproduce spectral distributions”, and “this effect is independent of the underlying architecture”.

- Dzanic et al. [12] take a step further and propose to exploit this frequency discrepancies for detection of deep network generated images, claiming an accuracy of up to 99.2% across multiple state-of-the-art GAN and VAE models.

Some works also propose different techniques to disguise these high frequency discrepancies via post-processing the deep network generated images [12, 25], or modifying the GAN training objective to avoid these discrepancies [10].
It should be noted that the cause of this discrepancy\textsuperscript{1} has not been agreed upon. Zhang et al.\textsuperscript{[44]} and Durall et al.\textsuperscript{[10]} suggest that this discrepancy could be caused by transposed convolution. As transposed convolution is used throughout the generator architectures, it is difficult to replace them. Therefore, Durall et al.\textsuperscript{[10]} propose spectral regularization to counteract this throughout the GAN training. Meanwhile, Dzanic et al.\textsuperscript{[12]} attribute this discrepancy to the linear dependencies in the spectrum of convolutional filters\textsuperscript{[25]}, which hinder learning of high frequencies.

\section*{1.2. CNN-generated Image Detectors}

Many works have addressed the possibilities of creating detectors to identify synthetic images apart from real images. Though synthetic image detectors (CNN-based) worked reasonably with RGB inputs\textsuperscript{[42, 5]}, with spectral discrepancies being observed, several works have proposed to use the corresponding Fourier representation to train these detectors\textsuperscript{[44, 13]}. In particular, Frank et al.\textsuperscript{[13]} showed that detectors using frequency domain yielded better results compared to using the RGB counterpart.

Though observations corresponding to distinguishable frequency footprints being left by CNN-generated images remained mostly qualitative, Dzanic et al.\textsuperscript{[12]} and Durall et al.\textsuperscript{[10, 11]} studied the spectral decay attributes and quantified this behavior via averaging the power spectrum over frequencies radially to represent them as 1D information. Through this, they observed high frequency spectral decay inconsistencies in CNN-generated images. Furthermore, Dzanic et al.\textsuperscript{[12]} proposed to use a simple KNN classifier using high frequency spectral attributes (3 features extracted per image) that surprisingly obtained very high accuracy in identifying synthetic images only with very small amount of training data. Similarly, Durall et al.\textsuperscript{[11]} used the entire 1D-power spectrum (contains all frequency information) as features to perform detection using Logistic regression, SVM and K-means clustering algorithms. These detectors\textsuperscript{[12, 11]} rely on the belief that high frequency decay discrepancies are intrinsic in CNN-generated images.

\section*{1.3. Our contributions}

In this work, we take a closer look at the high frequency decay discrepancies in CNN-generated images. Analysis of CNN-generated images is a daunting task: a large number of different architectures, algorithms and objective functions have been proposed to train generators. Instead, our study focuses on the last layer of the generators. Our justification is as follows. Based on Sampling Theorem\textsuperscript{[20]}, the frequency contents of the outputs of individual layers are limited by the sampling resolution of the corresponding outputs. As previous work has reported the discrepancies between real and CNN-generated images at the highest frequencies, we hypothesize that inner generator layers (which produce lower resolution outputs) are not directly responsible for the high frequency discrepancies. Therefore, we focus on the last upampling layer of generative CNN models.

Due to our focus being localized at the last layer, we are able to pinpoint the component that is related to this discrepancy across multiple GAN loss functions, architectures, datasets and resolutions. Our candidate GANs are similar to Durall et al.\textsuperscript{[10]}: DCGAN\textsuperscript{[34]}, LSGAN\textsuperscript{[31]}, WGAN-GP\textsuperscript{[17]} and we also extend to StarGAN\textsuperscript{[7]}. Our experiments suggest that the frequency discrepancies can be largely avoided by simply modifying the feature map scaling of the last layer. Importantly, using the same training algorithms, objective functions and network architectures (except using a different scaling in the last layer) as in standard GAN models, we are able to avoid the spectral dis-

\footnotetext{\textsuperscript{1}The terms spectral discrepancies and spectral inconsistency are used invariably where we refer to high frequency spectral decay discrepancies. We also use the terms CNN-generated and synthetic invariably.}
crepancies. Therefore, our work provides counterexamples to argue that high frequency discrepancies are not intrinsic for CNN-generated images. Furthermore, we are able to successfully bypass the synthetic image detector proposed by Dzanic et al. [12] with only such change in the last scaling step, showing that such approach may not be reliable for detection of deep network generated images. The key takeaway from our work is:

- High frequency spectral decay discrepancies are not intrinsic for CNN-generated images. Therefore, we urge re-thinking in using such features for CNN-generated image detection.

2. Related Work

Dzanic et al. [12] show that CNN-generated images (GANs and VAEs) demonstrate different Fourier spectrum decay characteristics. Since the spectra of natural images tend to behave following the power law [41], Dzanic et al. [12] show that the Fourier modes of deep network generated images at the highest frequencies did not decay as seen in real images, but instead stayed approximately constant. Furthermore, they propose to exploit these discrepancies to detect CNN-generated synthetic images by fitting a decay function to the reduced spectra, and using the parameters of the fitted decay function to build a simple kNN classifier.

Durrall et al. [10] show that popular GAN image generators fail to approximate the spectral distributions of real data, and they attribute this to the use of transpose convolutions for upsampling. They show that this effect is independent of the underlying architecture using 1-dimensional spectral characteristics of images generated from DCGAN [34], LSGAN [31], WGAN-GP[17] and DRAGAN [26]. Since transpose convolutions are used in the entire generator models, the propose to counteract their effects by adding a spectral regularization term to the Generator, thereby penalizing the generator for spectral distorted samples.

Khayatkhoei and Elgammal [25] suggest the presence of a systematic bias in GANs against learning high frequencies. They specifically show that for a given kernel size, as resolution increases, the correlation/dependency of the kernel’s spectrum increases thereby systematically preventing GANs to learn high frequencies without affecting the adjacent frequencies. To alleviate this shortcoming, they propose frequency shifted generators whose frequencies are shifted towards specific high frequencies.

3. Background

The 2D discrete Fourier transform $F(k_x, k_y)$ of a $M \times N$ 2D image $f(x, y)$ can be written as,

$$F(k_x, k_y) = \frac{1}{MN} \sum_{u=0}^{M-1} \sum_{v=0}^{N-1} f(u, v) e^{-j2\pi (\frac{k_x u}{M} + \frac{k_y v}{N})}$$  \hspace{1cm} (1)$$for $k_x \in \{0, 1, 2, ..., M - 1\}$, $k_y \in \{0, 1, 2, ..., N - 1\}$.

We follow the convention in [12, 10] and compute the azimuthally average of the magnitude of Fourier coefficients over radial frequencies to obtain the reduced spectrum and normalize it. The reduced spectra indicates the strength of the signal with respect to different spatial frequencies. Since our study focuses on high frequency Fourier attributes, we pay attention to the last 25% of the spatial frequencies (0.75 - 1.0 normalized spatial frequencies) similar to [12].

The Sampling Theorem ([20], chapter 4.2) states that: A bandlimited image $f(x, y)$ with bandwidths $\xi_{x0}, \xi_{y0}$ can be recovered without error from the sample values provided the sampling rate is greater than the Nyquist rate: $2\xi_{x0}, 2\xi_{y0}$. The image $f(x, y)$ with bandwidths $\xi_{x0}, \xi_{y0}$ means that there is no frequency content outside a bounded region in the frequency plane defined by $\xi_{x0}, \xi_{y0}$. Details and mathematical proofs can be found in [20] chapter 4.2.

4. Last Upsampling Operation and Fourier Discrepancies

Based on last section, the maximum frequency represented by a discrete 2D signal is constrained by the spatial resolution (sampling) of the signal. Previous work has consistently reported discrepancies in the highest frequencies [10, 12]. Therefore, we hypothesize that inner generator layers that produce lower resolution outputs may not be directly responsible for the high frequency discrepancies. Therefore, we focus on the last upsampling step. In particular, we split the last step into 2 operations, 1) Feature map scaling and 2) Convolution.

4.1. Feature Map Scaling

Feature map scaling is a non-parametric operation that scales the input in both dimensions by some factor (Usually 2 in most GAN architectures). In this work, we focus on 3 common feature map scaling techniques. 1) Zero-insert scaling inserts zero between every row and column, scaling the input in both dimensions which is also used by transpose convolutions. 2) Nearest interpolation scales the input by inserting nearest neighbour values. 3) Bilinear interpolation scales the input by inserts new values by taking the weighted average of adjacent values. From frequency perspective, zero-insertion introduces the largest amount of high frequency content as it replicates the low frequency spectrum for the high frequencies [44], followed by bilinear and nearest interpolation. We focus more on the “spectral trend” than the frequency values, and show a schematic example of these upsampling effects on the normalized reduced spectra in Figure 2 by upsampling a reference image of 128x128 from CelebA. [29]
4.2. Convolution

The subsequent convolution operation learns kernels in order to satisfy the optimization objective. Convolutional kernels are capable of suppressing/amplifying high frequencies. e.g. A Gaussian kernel suppresses high frequencies and a bilateral kernel amplifies high frequencies. So when designing upsampling blocks in GANs, the general intuition is that irrespective of the feature map scaling method, the kernels will learn to manipulate the scaled feature maps to satisfy the objective function.

5. Experiments

Here we discuss the main experiments. Additional experiments and analysis can be found in Supplementary.

In order to investigate the effects of these 2 operations, we design a rigorous test bed that can address feature map scaling, kernel size and number of kernels independently. Our aim is to isolate the effects of these 2 operations and understand their roles (if any) in causing the high frequencies discrepancies. We use celebA [29] dataset at 128x128 resolution and use 3 GANs with identical architectures but different loss functions namely 1) DCGAN [34], 2) LSGAN [31] and 3) WGAN-GP [17]. All baseline models consist of transpose convolutions with kernel size 4 identical to most of the box GAN architectures including CycleGAN [46], StarGAN [7] and VQ-VAE [40].

**Proposed Test Bed.** Table 1 summarizes our test bed. All baseline experiments using transpose convolutions of kernel size 4 are indicated by the experiment code Baseline and our handcrafted experiments are indicated using the 3 character code: An experiment code of X.Y.Z indicates X type of feature map scaling (Possible values are Z: Zero-insertion, N: Nearest interpolation, B: Bilinear interpolation), Y number of convolutional blocks and Z sized convolutional kernels for the last upsampling step. e.g. A code of N.1.5 indicates nearest interpolation feature map scaling with a single convolutional block of 5x5 kernel for the last upsampling step. Note that we focus on up scaling by a factor of 2 as used in most GAN models. Our test bed summary is shown in table 1 and it contains experiments addressing the following factors:

**Feature Map scaling.** In order to investigate the effect of feature map scaling on high frequency Fourier attributes, we use experiments Z.1.5, N.1.5 and B.1.5. We explicitly conduct experiments using zero-insert feature map scaling as sanity check experiments to verify the effects of transpose convolutions. Do note that we had to use odd size kernel to maintain the spatial size after scaling, and hence we use kernels of size 5 for the last convolutional block.

**Kernel Size.** We use experiments x.1.3, x.1.5 and x.1.7 to investigate the kernel size effect on high frequency Fourier behaviour. Here x refers to Z, N or B for different types of scaling as discussed.

**Number of kernels.** Further, we use experiments x.1.5 and x.3.5 to study the effect of number of kernels on the high frequency Fourier behaviour.

6. Metrics

Spectral behaviour of synthetic images have been analysed only qualitatively, not quantitatively in previous works [12, 10, 11]. Defining a spectral consistency metric is non-trivial and to be consistent, we will qualitatively analyse the high frequency spectral distribution in our experiments. Synthetic images are spectral consistent if they demonstrate power spectrum decay behaviour similar to their training data, and if not, vice versa. We use 4000 real and GAN images to generate spectral distribution curves. To ensure
that our setups were trained properly, we use FID scores to assess the quality of samples and make sure that they are consistent with the FID scores from baseline experiments.

7. Results

7.1. Effect of Feature Map Scaling Methods

Figure 3 illustrates the resulting spectral distributions when using different feature map scaling methods. We observe that images from Z.1.5 and Baseline experiments are spectral inconsistent for all 3 GANs. Nearest and bilinear interpolation methods are able to replicate the spectral distribution of real data reasonably across all 3 GAN models. Since the only change in our models is the feature map scaling method in the last layer while the algorithms, objective functions and majority of the network architectures are identical to public released code, these results qualify to support our thesis that high frequency Fourier discrepancies are not inherent to GANs.

7.2. Effect of Kernel Size

From Figure 4, we observe that smaller kernel sizes result in more turbulent spectral behaviour for N.1.3 and B.1.3 LSGAN experiments. Apart from this observation, experiments using nearest and bilinear interpolation methods are able to reproduce spectral distributions of real data reasonably well, and zero-insertion based methods always result in high frequency spectral distortions. Further, from Figure 5, we observe that when using larger kernels, all experiments using nearest and bilinear interpolation methods are able to replicate spectral behaviour of real data, and even with larger kernels zero-insertion based methods produce high frequency spectral distortions (Z.1.7).

7.3. Effect of Number of Kernels

Figure 6 illustrates that increasing the number of kernels do not yield spectral consistency by itself. Apart from N.3.5 DCGAN experiment, we observe that all other experiments using nearest and bilinear interpolation methods are able to approximate the spectral behaviour of real data. Also, GAN objective functions do not impose any spectral requirements. Thus the kernels do not have any direct incentive for imposing spectral consistency.

Key observations. Throughout all 39 rigorous experiments, we observe that zero insertion based feature map scaling methods (including Baseline) are consistently showing high frequency spectral discrepancies, and most experiments (22/24) using bilinear and nearest interpolation methods are able to avoid these high frequency spectral discrepancies. All these results support our statement that high frequency spectral discrepancies are not inherent characteristics to GANs. The FID scores for all experiments were comparable with the baseline FID, and are included in the supplementary. We show image samples from WGAN-GP for Baseline, N.1.5 and B.1.5 setups in Figure 7. Do note that for all experiments, we use the exact same discriminator architecture as the Baseline experiment.

8. Fourier Synthetic Image Detector

Dzanic et al. [12] state that the spectral properties of real and deep network generated images are fundamentally different, and proposed a synthetic image detection method using a “simple” k-nearest neighbours (KNN) classifier to emphasize the extent of these spectral differences. We follow the exact procedure as the original authors to train these classifiers and details can be found in Supplementary.

8.1. Experiment Setup

With nearest and bilinear interpolation methods obtaining spectral consistent GANs for the previous experiments, we question whether the classifier proposed above would be robust enough to detect these samples as fake. To investigate this we follow the following steps:

1. We train 3 KNN classifiers, one for DCGAN, LSGAN and WGAN-GP respectively. For GAN images, we use images generated from the Baseline experiment (using transpose convolutions) as training data.

2. We test these classifiers using GAN images generated from the setups in our test bed to evaluate the robustness of the classifier.

3. We also repeat the experiments using 50% data for training the classifier (The original work used only 10%) to observe any improvements in accuracy.

8.2. Detection Results

The complete detection results are shown in table 2. We observe that all setups corresponding to N.x.x and B.x.x experiments are able to easily bypass the classifier. Even when 50% training data is used, we are able to bypass the classifier with ease (Included in Supplementary). The results clearly demonstrate that the proposed classifier relying on high frequency Fourier attributes to detect synthetic images, fails to detect images generated from identical GAN models with last feature map scaling replaced by nearest or bilinear interpolation methods. These results are consistent with the observed spectral distributions. By combining these detection results with the empirical finding that high frequency spectral discrepancies are not inherent characteristics of CNN-generated images, we suggest re-thinking of using such discrepancies to detect synthetic images.

9. Extended experiments

With observations that high frequency Fourier spectrum discrepancies are not intrinsic characteristics of GANs, we
extend our experiments to address different dataset, image resolution and GAN objective function to further find evidences to support our thesis statement. We select 3 setups from our test bed Z.1.5, N.1.5 and B.1.5 to conduct extended experiments since we have observed that kernel size/number of kernels do not substantially manipulate high frequencies compared to feature map scaling methods. Similar to previous experiments, we analyze the resulting spectral distributions and evaluate the robustness of the synthetic image detector proposed by Dzanic et al. [12].

### 9.1. LSUN Bedrooms Dataset

In this experiment we use a subset of LSUN Bedrooms Dataset [43] (128x128) to train DCGAN [34], LSGAN [31] and WGAN-GP [17] identical to previous setups. The spectral plots are shown in Figure 8. We observe identical results to CelebA experiment (Figure 3). That is we observe N.1.5 and B.1.5 are producing spectral consistent GANs and this further supports our statement that high frequency spectral discrepancies are not inherent in GANs. We also evaluate the synthetic image detector and observe that N.1.5 and B.1.5 samples can easily bypass the detector. (See table 3)

### 9.2. Image-to-Image Translation

We extend our experiments to Image-to-Image translation domain using StarGAN [7]. Here we use resized CelebA [29] (256x256) used by the official StarGAN [7] implementation to study whether spectral consistency can be achieved by modifying the last feature map scaling oper-
The recent Spectral regularization (SR) by Durall et al. [10] proposed to add a regularizer term to the generator loss to explicitly penalize the generator for spectral distortions. Using SR, they were able to obtain spectral consistency for DCGAN [34], LSGAN [31], WGAN-GP [17] and DRAGAN [26] using the celebA [29] (128x128). This method encounters computational overhead due to calculation of reduced spectrum for images during training. We show that by modifying the last feature map scaling operation, we are able to achieve spectral consistent GANs without SR. Importantly, we show that no change in objective functions is needed. These results and more discussion on SR can be found in Supplementary.

11. Discussion

In this study, we investigated the validity of contemporary beliefs that CNN-based generative models are unable to reproduce high frequency decay attributes of real images. We employ a systematic study to design counterexamples to challenge the existing beliefs. With maximum frequency bounded by the spatial resolution, and Fourier discrepancies reported at the highest frequencies, we hypothesized that the last upsampling operation is mostly related to this shortcoming. With carefully designed experiments span-
Figure 7. WGAN-GP samples for Baseline (Left), N.1.5 (Middle) and B.1.5 (Right) for CelebA [29]. We observe that the visual quality is comparable when replacing the last transpose convolutions with nearest and bilinear methods. More visual results in supplementary.

Figure 8. LSUN results. We observe spectral plots identical to CelebA [29] experiments. Refer to table 1 for experiment details.

Figure 9. Spectral plots for StarGAN images. We observe that Bilinear feature map scaling produces spectral consistent GANs.

Acknowledgments: This project was supported by SUTD project PIE-SGP-AI-2018-01. This research was also supported by the National Research Foundation Singapore under its AI Singapore Programme [Award Number: AISG-100E2018-005]. This work was also supported by ST Electronics and the National Research Foundation (NRF), Prime Minister’s Office, Singapore under Corporate Laboratory @ University Scheme (Programme Title: STEE Infosec - SUTD Corporate Laboratory).

We also gratefully acknowledge the support of NVIDIA AI Technology Center (NVATIC) for our research.

In Supplementary material, we provide more GAN models [6, 28] with no high frequency decay discrepancies. We also investigate whether such high frequency decay discrepancies are found in other types of computational image synthesis methods (synthesis using Unity game engine\(^2\)) [15, 14]. To conclude, through this work we hope to help image forensics research manoeuvre in more plausible directions to combat the fight against CNN-synthesized visual disinformation.

\(^2\)https://unity.com/
References


[30] Sophie Maddocks. ‘a deepfake porn plot intended to silence me’: exploring continuities between pornographic and ‘political’ deep fakes. Porn Studies, 0(0):1–9, 2020. 1


[32] Rachel Metz. The number of deepfake videos online is spiking. most are porn, Oct 2019. 1


[36] Tom Simonite. What happened to the deepfake threat to the election?, Nov 2020. 1

[37] Daniel Thomas. Deepfakes: A threat to democracy or just a bit of fun?, Jan 2020. 1


[42] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. Cnn-generated images are surprisingly easy to spot... for now. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020. 1, 2


