Spatial Frequency Bias in Convolutional Generative Adversarial Networks

Mahyar Khayatkhoei, Ahmed Elgammal

Department of Computer Science, Rutgers University
New Brunswick, New Jersey
{m.khayatkhoei, elgammal}@cs.rutgers.edu

Abstract

Understanding the capability of Generative Adversarial Networks (GANs) in learning the full spectrum of spatial frequencies, that is, beyond the low-frequency dominant spectrum of natural images, is critical for assessing the reliability of GAN-generated data in any detail-sensitive application. In this work, we show that the ability of convolutional GANs to learn an image distribution depends on the spatial frequency of the underlying carrier signal, that is, they have a bias against learning high spatial frequencies. Our findings are consistent with the recent observations of high-frequency artifacts in GAN-generated images, but further suggest that such artifacts are the consequence of an underlying bias. We also provide a theoretical explanation for this bias as the manifestation of linear dependencies present in the spectrum of filters of a typical generative Convolutional Neural Network (CNN). Finally, by proposing a proof-of-concept method that can effectively manipulate this bias towards other spatial frequencies, we show that the bias is not fixed and can be exploited to explicitly direct computational resources towards any specific spatial frequency of interest in a dataset, with minimal computational overhead.

1 Introduction

The information contained in an image is carried by a set of spatial frequencies, that is, a set of planar sinusoids with unique frequencies and directions. Intuitively, we associate the high frequencies with the details of the image, and the low frequencies with its general form; however, neither frequencies should be treated as more important by a generative model seeking to learn a distribution. To make this more clear, consider a two-dimensional planar cosine wave defined over a 128×128 image, and assume that we sample the magnitude of this static wave from a Gaussian distribution. Whether this wave completes 64 periods across the image (*i.e*. high frequency), or 3 periods (*i.e*. low frequency), ideally should not affect a generative model's learning of the underlying Gaussian distribution (see Appendix [1] for an empirical realization of this thought experiment).

Convolutional Generative Advarsarial Networks (GANs) (Goodfellow et al. 2014; Radford, Metz, and Chintala 2015) are the foremost generative models for generating image distributions, and while many of their limitations have been studied from the perspective of probability theory and manifold learning (Arjovsky and Bottou 2017; Arora et al. 2017; Khayatkhoei, Singh, and Elgammal 2018), their spectral limitations remain understudied. Importantly, the theory of GANs (Goodfellow et al. 2014), and its many variants, do not suggest any spectral limitation. Nevertheless, the progression of GAN research over the recent years reflects a constant effort for generating better *details* while generating *general form and color* seems to be quite easy (Karras et al. 2020; Karras, Laine, and Aila 2019; Karras et al. 2018; Wang et al. 2018; Huang et al. 2017). In this work, we investigate whether this difficulty can be linked to a spatial frequency bias. In particular, we find that high and low spatial frequencies are not treated equally by convolutional GANs, and that the information carried by high frequencies are more prone to loss. Our findings imply that when convolutional GANs are used as part of any detail-sensitive application, for example in augmenting or correcting medical and satellite images, generated fine details are not as reliable as the overall shape and form.

Very relevant to this work, two recent concurrent works (Dzanic, Shah, and Witherden 2020; Durall, Keuper, and Keuper 2020) have observed that high-frequency discrepancies can be utilized to easily distinguish GAN generated images from natural images. Specifically, Dzanic *et al*. (2020) have shown that these discrepancies can be removed by directly modifying the spectrum of generated images in post-processing. Our findings in this work are consistent with these observations, but further suggest that such discrepancies are the consequence of an underlying spatial frequency bias against learning high frequencies. As such, the issue is not merely that GANs generate high-frequency artifacts on datasets with little to no high-frequency content (*e.g*. high resolution natural images), which can be removed by modifying the generated spectrum after training, rather that GANs tend to lose the information carried by high frequencies, which is not recoverable in post-processing. The main findings and contributions of this work are listed below:

- Convolutional GANs trained on natural images do not learn high spatial frequencies as well as low spatial frequencies, suggesting the existence of a spatial frequency bias (Section 3.1).

[1] Extended version available at https://arxiv.org/abs/2010.01473

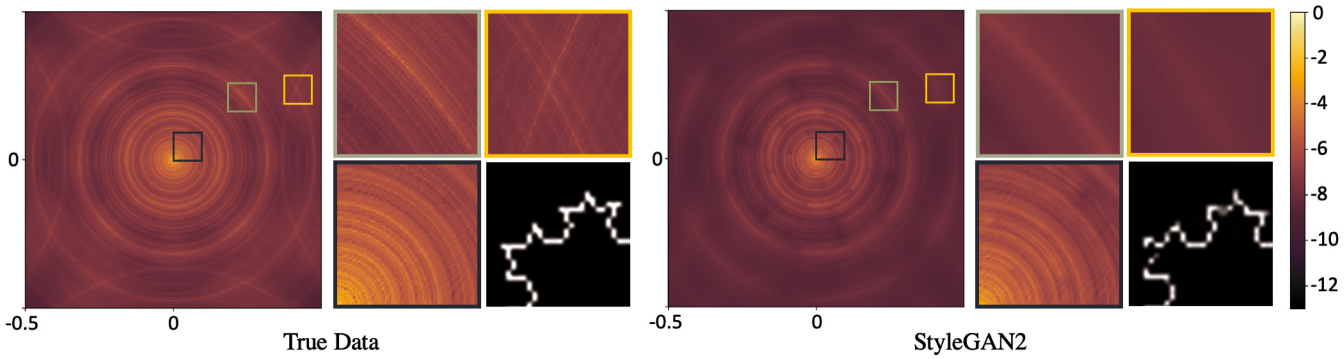

Figure 1: Average power spectrum of a large-scale GAN trained on a fractal-based dataset clearly reveals how the low frequencies (closer to center) are matched much more accurately than the high frequencies (closer to corners). (Left) Average power spectrum of randomly rotated Koch snowflakes of level 5 and size 1024×1024. (Right) Average power spectrum of StyleGAN2 trained on the latter. A representative patch from the perimeter of true and generated fractals are also displayed.

- The same distribution when primarily carried by high spatial frequencies becomes harder to learn for convolutional GANs, versus when carried by low spatial frequencies, confirming the bias (Section 3.2).
- The bias is theoretically explained as a manifestation of linear dependencies contained in the spectrum of filters of a generative Convolutional Neural Network (CNN) (Section 3.3).
- A proof-of-concept method is proposed to show that the bias is not fixed and can be efficiently shifted towards other spatial frequencies (Section 4).

Note that in this work, we are only considering GANs that use CNNs as their generator, and for brevity, we refer to these models simply as GANs. We use three popular convolutional GAN models in our studies: WGAN-GP (Gulrajani et al. 2017) serves as a simple but fundamental GAN model; and Progressively Growing GAN (PG-GAN) (Karras et al. 2018) and StyleGAN2 (Karras et al. 2020) serve as state-of-the-art models with large capacity and complex structure, incorporating state-of-the-art normalization and regularization techniques. Since our goal is to compare the performance of GANs on high versus low spatial frequencies, and not to compare the overall quality of the generated samples with one another or the state-of-the-art, we chose to use PG-GAN and StyleGAN2 with a slightly smaller capacity in our training (corresponding to the capacity used in Section 6.1 of Karras et al. (2018) for ablation studies). See Appendix for the details of each model.

2 Spatial Frequency Components

According to Inverse Discrete Fourier Transform, every periodic discrete 2D signal $I(x, y)$ with $x \in \{0, 1, 2, \dots, m-1\}$ and $y \in \{0, 1, 2, \dots n-1\}$, can be written as a sum of several complex sinusoids as follows:

$$\begin{aligned} I(x,y) &= \frac{1}{mn} \sum_{u=0}^{m-1} \sum_{v=0}^{n-1} C(u,v) e^{j2\pi(\frac{ux}{m} + \frac{vy}{n})} \\ &= \frac{1}{mn} \sum_{u=0}^{m-1} \sum_{v=0}^{n-1} C(u,v) e^{j2\pi(\hat{u},\hat{v}).(x,y)} \end{aligned} \tag{1}$$

We denote each complex sinusoid a *spatial frequency component* which can be expressed by a vector (u, v) over the pixel locations (x, y). In the above equation, $C(u, v)$ is the complex amplitude of each frequency component, $(\hat{u}, \hat{v}) = (\frac{u}{m}, \frac{v}{n})$ defines the direction of propagation on the 2D plane and its magnitude defines the spatial frequency in that direction, and $m, n \in \mathbb{N}$ are the periods of I in x and y direction respectively. Every channel of a digital 2D image can be assumed periodic beyond the image boundaries, and therefore represented by Eq. (1), with periods m and n being the length and width of the image respectively. In that case, the vector $(\hat{u}, \hat{v})$ would define the spatial frequency of a sinusoid in terms of *cycles per pixel*, in x and y direction respectively, with $\hat{u}, \hat{v} \in [0, 1)$. The maximum frequency in each direction is 0.5 corresponding to the Nyquist frequency (the shortest period needs at least two pixels to be represented, hence the maximum frequency is half cycle per pixel). In favor of clarity, and without loss of generality, we will assume $\hat{u}, \hat{v} \in [-0.5, 0.5)$ throughout this paper. Additionally, we loosely refer to the spatial frequency components with $|\hat{u}|$ or $|\hat{v}|$ close to 0.5 as high frequencies, and with $\hat{u}$ or $\hat{v}$ close to 0 as low frequencies. Whenever displaying power spectrums $|C(\hat{u}, \hat{v})|^2$, for better visualization, we drop the DC power, apply Hann window, normalize by the maximum power, and apply $\log$, such that the most powerful frequency always has value 0. Also, $\hat{u}$ and $\hat{v}$ are placed on horizontal and vertical axes respectively, such that low frequencies are placed close to the center, while high frequencies close to the corners.

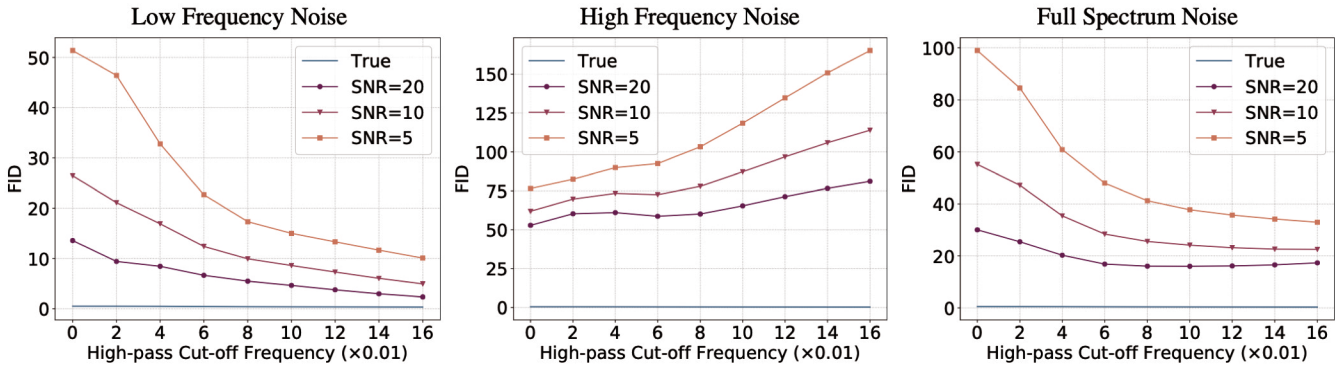

Figure 2: Sensitivity of FID Levels to the presence of mismatch in different spatial frequency bands. FID Levels between two disjoint subset of CelebA images are plotted after adding spectral noise to one of the sets, at a low-frequency band $[0, \frac{1}{8}]$ (left), a high-frequency band $[\frac{3}{8}, \frac{1}{2}]$ (middle), and all frequencies (right). The *blue curve* depicts True FID Levels (no added noise).

3 The Spatial Frequency Bias

3.1 FID Levels

We first want to observe whether GANs trained on natural images learn the information carried by high frequencies as well as low frequencies. One approach is to directly compare the average power spectrums of GAN generated images with true images (Dzanic, Shah, and Witherden 2020; Durall, Keuper, and Keuper 2020). However, the average power of a spatial frequency component is not very informative of how well the distribution carried by that component is learnt. A more informative alternative is to compare how well the distribution of image features carried by high frequencies are learnt compared to that carried by low frequencies. Frechet Inception Distance (FID) (Heusel et al. 2017) provides a reliable measure of mismatch between the distributions of features extracted from two sets of images: the larger the FID, the larger the mismatch. We propose an extension to this metric, denoted *FID Levels*, in which we plot FID between two sets of images after gradually removing low frequency bands from both sets. Each point on the FID Levels plot shows FID computed after applying a high-pass Gaussian filter, with the cut-off specified on the horizontal axis, to both the GAN generated and the true images (one standard deviation in the Fourier domain is considered as filter cut-off). As a baseline for comparison, we also compute FID Levels between two disjoint subsets of the true images, denoted *True FID Levels*. For completeness, and as a direct measure of spectral difference, we also report total variation (Gibbs and Su 2002) between the GAN generated and the true average power spectrums normalized into density functions, denoted *Leakage Ratio (LR)*.

If a generative model is learning low and high frequencies equally well, then gradually removing spatial frequency bands from generated and true images should result in a generally declining FID between the two sets, as the total amount of information is gradually reduced. To illustrate this behavior, we consider a noisy version of True FID Levels, where we manually introduce mismatch at different spatial frequencies by perturbing the frequencies of one of the two disjoint sets of true images. Specifically, we perturb a frequency component by adding normal noise with mean 0 and variance equal to its power. The total noise added to each image is normalized such that a fixed signal to noise power ratio (SNR) is maintained. Figure 2 shows that when a fixed amount of noise (in terms of SNR) is introduced at low frequencies, or at all frequencies, the FID Levels declines; in contrast, when the same amount of noise is introduced at high frequencies, the FID Levels increases.

Given the observations in Figure 2, we return to the case of GANs. Figure 3 shows FID Levels of GANs trained on two 128×128 image datasets: CelebA (Liu et al. 2015) and LSUN-Bedrooms (Yu et al. 2015). The GANs exhibit an increase in FID Levels on both datasets, similar to the behavior observed in Figure 2 (middle), suggesting that the learnt high frequencies contain more mismatch than the low frequencies. In contrast, the True FID Levels remains approximately constant in both datasets. As such, the GANs appear to have a bias against learning high frequencies. Without such a bias, we would expect a declining FID Levels as previously observed in Figure 2 (right). There remains a caveat, however; since much of the information in natural images is concentrated at low frequencies (Field 1987), this bias could be attributed to the scarcity of high frequencies during training. We will see that this cannot be a sufficient explanation in the next subsection.

3.2 High-Frequency Datasets

If GANs are not biased against high frequencies, their performance should remain indifferent to shifting the frequency contents of the datasets. In other words, whether the information in a dataset is primarily carried by high frequencies or low frequencies should not affect how well GANs can learn the underlying image distribution. In order to test this hypothesis, we can create high-frequency versions of CelebA and LSUN-Bedrooms by applying a frequency shift operator, that is, multiplying every image in each dataset with $\cos(\pi(x + y))$ channel-wise, to create Shifted CelebA (SCelebA) and Shifted LSUN-Bedrooms (SBedrooms) respectively. In effect, all we are doing is swapping the low and high frequency contents of these datasets. Note that the

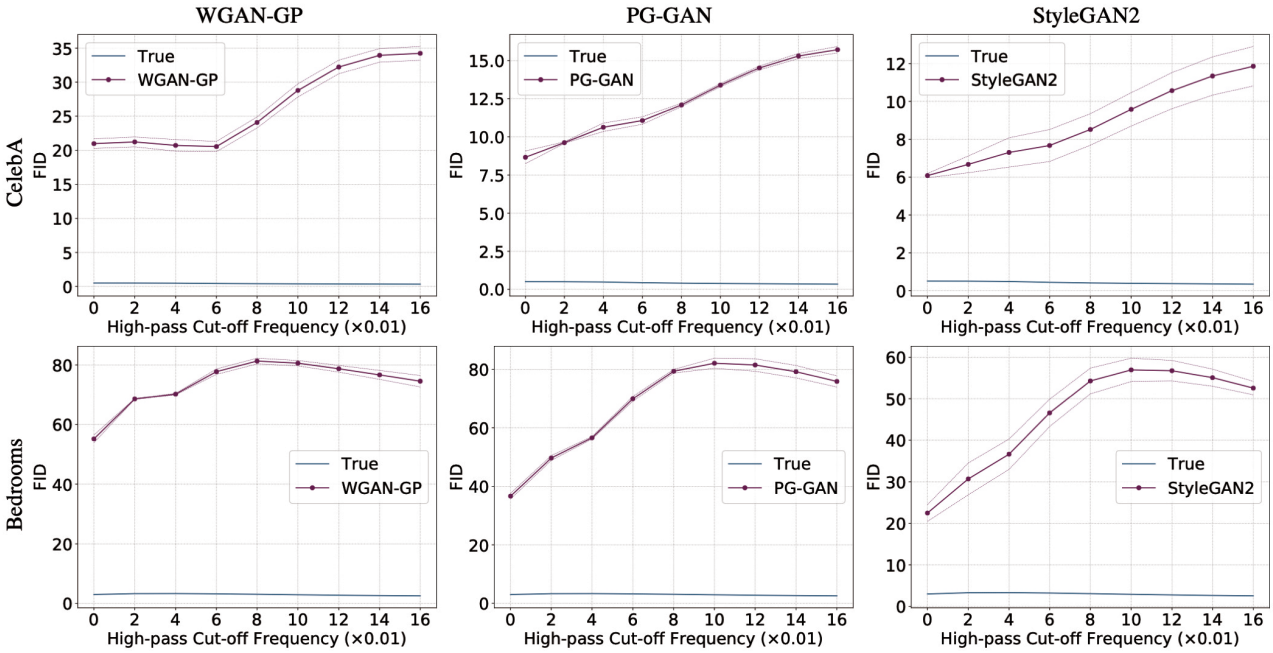

Figure 3: FID Levels of GANs trained on CelebA and LSUN-Bedrooms. The farther to the right on the horizontal axis, the more low frequencies are removed prior to FID computation. Notice the transient increase in FID (worsening performance) as low frequencies are removed. In all figures, the *blue curve* depicts the True FID Levels of the corresponding dataset as a baseline. All figures show average FID with one standard deviation error bars (*dashed line*), over three random training runs.

Model	Score	CelebA	SCelebA	Bedrooms	SBedrooms
WGAN-GP	FID	20.97 ± 0.70	328.72 ± 9.70	55.14 ± 1.29	283.02 ± 7.06
	LR (%)	2.29 ± 0.31	59.04 ± 5.09	1.99 ± 0.48	42.42 ± 4.32
PG-GAN	FID	8.66 ± 0.41	23.12 ± 2.08	36.65 ± 0.97	69.03 ± 10.28
	LR (%)	1.06 ± 0.21	3.93 ± 0.70	1.51 ± 0.29	3.12 ± 0.16
StyleGAN2	FID	6.08 ± 0.11	343.57 ± 53.59	22.49 ± 2.00	260.84 ± 4.03
	LR (%)	1.55 ± 0.42	43.03 ± 34.10	1.28 ± 0.19	7.32 ± 1.86

Table 1: Performance drop (increase in FID and LR) in GANs trained on the high-frequency versions of CelebA and LSUN-Bedrooms (SCelebA and SBedrooms). Average with one standard deviation ($\pm$) is reported over three random training runs.

frequency shift operator is a homeomorphism and therefore the distributions of SCelebA and SBedrooms have the same topological properties as CelebA's and LSUN-Bedroom's, and therefore the GANs' performance should remain unchanged from a purely probabilistic perspective.

Table 1 compares the GANs' performance on SCelebA and SBedrooms versus the original CelebA and LSUN-Bedrooms.[2] Their performance has worsened significantly (larger FID and LR) on the high-frequency datasets, showing that the GANs perform considerably better when the same image distribution is carried primarily by low frequencies. This observation rejects our earlier hypothesis in the start of this subsection, and confirms that the GANs' performance is sensitive to frequency shift. Additionally, this shows that the bias against high frequencies we observed in Section 3.1 cannot be explained by the scarcity of high frequencies in natural images: even though the unbalancedness in the distribution of power has remained unchanged in the high-frequency versions of the datasets, the GANs' performance has worsened significantly. We conclude that this bias is indeed a spatial frequency bias against high frequencies, regardless of how abundant or scarce they are in the dataset.

[2]In SCelebA and SBedrooms, true and generated images are re-shifted before computing FID so that the values are comparable with the FID results on CelebA and LSUN-Bedrooms.

3.3 On the Cause of the Bias

The three GAN models in which we observed the spatial frequency bias each have a unique network structure, and various aspects of their respective structures can cause or affect the bias, as evident from the difference in their performance on high-frequency datasets in Table 1; nonetheless, a

common aspect of all these models is the use of generative CNNs, and in this subsection we will theoretically expose a limitation in the generative CNN that can play a fundamental role in causing the bias.

Let us consider the structure of a typical generative CNN. A 2D generative CNN $G(x, y; W, H^1)$, with parameters $W \in \mathcal{W}$, input features $H^1 \in \mathbb{R}^{d_0 \times d_0 \times c_0}$, and output space $\mathbb{R}^{d \times d}$, can be modeled as a series of affine convolution layers $\text{Conv}_i^l : \mathbb{R}^{d_{l-1} \times d_{l-1} \times c_{l-1}} \rightarrow \mathbb{R}^{d_l \times d_l \times c_l}$ as follows: [3]

$$H_i^{l+1} = \text{Conv}_i^l(H^l) = b_i + \sum_c F_{ic}^l * \text{Up}\left(\sigma(H_c^l)\right) \quad (2)$$

where l indices the layer (depth), i the output channels (width), c the input channels, $F_{ic}^l \in \mathbb{R}^{k_l \times k_l}$ is a parametric 2D filter, $b_i \in \mathbb{R}$ is the bias, $\text{Up}(.)$ denotes the upsampling operator, and $\sigma(.)$ is a non-linearity. If we restrict σ to rectified linear units (ReLU), then in a neighborhood of almost any parameter W, we can consider the combined effect of $\text{Up}(\sigma(.))$ as a fixed linear operation:

Proposition 1. *At any latent input H^1 of a finite size ReLU-CNN, almost everywhere on the parameter space, there exists a neighborhood in which ReLUs are equivalent to fixed binary masks.* [4]

Therefore, in this neighborhood, improving the output spectrum is only achievable through adjusting the spectrum of filters F_{ic}^l. Intuitively, the filters try to *carve out* the desired spectrum out of the input spectrum which is distorted by ReLUs (as binary masks), and aliased by upsampling. In the following theorem, we investigate how freely these filters can adjust their spectrum. Specifically, we will show how the filter size k_l and the spatial dimension d_l of a Conv layer affect the correlation in the spectrum of its filters. Note that more correlation in a filter's spectrum means more linear dependency, and thus reduces its effective capacity, in other words, the filter can not freely adjust specific frequencies without affecting the adjacent correlated frequencies.

Theorem 1. *Let $U = \mathcal{F}\left\{F_{ic}^l\right\}(u_0, v_0)$ and $V = \mathcal{F}\left\{F_{ic}^l\right\}(u_1, v_1)$ be any two spatial frequency components on the spectrum of any 2D filter of the l-th Conv layer, with spatial dimension $d_l \in \mathbb{N}$ and filter size $k_l \in \mathbb{N}$, such that $1 < k_l \leq d_l$. Assuming i.i.d. weight initialization, the magnitude of the complex correlation coefficient between U and V, at any point during the training, is given by:* [4]

$$\begin{aligned} |corr(U,V)| &= \left| \frac{Sinc(u_0 - u_1, v_0 - v_1)}{k_l^2} \right| \\ s.t. \quad Sinc(u,v) &= \frac{\sin(\frac{\pi u k_l}{d_l}) \sin(\frac{\pi v k_l}{d_l})}{\sin(\frac{\pi u}{d_l}) \sin(\frac{\pi v}{d_l})} \end{aligned} \quad (3)$$

Corollary 1.1. *If U and V are two diagonally adjacent spatial frequency components of F_{ic}^l, then:*

$$|corr(U,V)| = \frac{\sin^2(\frac{\pi k_l}{d_l})}{k_l^2 \sin^2(\frac{\pi}{d_l})} \quad (4)$$

[3] Transposed Conv layers are sufficiently represented by an appropriate choice of the $\text{Up}(.)$ operator.

[4] Proof in Appendix.

Now, in each Conv layer, note that the maximum spatial frequency that can be generated is limited by the Nyquist frequency, that is, Conv^l can only adjust image spatial frequencies in $[0, \frac{d_l}{2d}]$ without aliasing[5]. This means that high frequencies are primarily generated by the CNN's outer layers where d_l is larger. According to Eq. (4), given a fixed filter size k_l, the larger the d_l, the larger the correlation in the filter's spectrum (see Appendix for the graph of this function), and consequently the smaller its effective capacity. Therefore, the outer layers responsible for generating high frequencies are more restricted in their spectrum compared to the inner layers with smaller d_l. We hypothesize that this is the main cause of the spatial frequency bias. This in turn implies that the issue of spatial frequency bias is not limited to GANs. Indeed, high-frequency discrepancies between CNN generated images and true images have been observed both in L2 reconstructions (Deng et al. 2020; Li and Barbastathis 2018; Ulyanov, Vedaldi, and Lempitsky 2018) and Variational Autoencoders (VAEs) (Dzanic, Shah, and Witherden 2020), however, in these tasks, the spatial frequency bias of generative CNNs is not easily distinguishable from the known spectral biases inherent to the respective objective functions.

On the effect of increasing depth. One way to counter the correlation is to replace an individual Conv layer with a stack of Conv layers, resulting in a deeper CNN. This can increase the effective filter size k_l operating on each spatial dimension, thus reducing the correlation. However, note that while only outer layers can generate high frequencies without aliasing, low frequencies can be generated by all layers without aliasing. As such, low frequencies will always enjoy a larger end-to-end filter size compared to high frequencies, and thus less correlation (see Appendix for visualization of the spectra of effective filters in trained WGAN-GP).

On the effect of increasing width. Another way to counter the correlation is to simply include more filters in a Conv layer, resulting in a wider CNN. However, this becomes particularly costly at the outer layers with larger spatial dimensions. Moreover, making the outer layers wider will increase the capacity of generating both high and low frequencies equally, and not exclusively that of high frequencies, as discussed in the previous paragraph, thus, the spatial frequency bias remains.

On the effect of increasing resolution. It is key to note that whether a signal contains high spatial frequencies or not is directly related to its sampling rate. For example, consider the continuous-valued image of a bird formed on a camera's sensor, whose feathers change color from white to black 64 times over the length of the image. If this image is sampled into a 128×128 picture, the feathers would form a high-frequency component ($\frac{1}{2}$ cycles per pixel). If the same image is instead sampled into a 1024×1024 picture, the feathers now form a low-frequency component ($\frac{1}{16}$ cycles per pixel). Therefore, one solution to the spatial frequency bias is to

[5] Aliasing here refers to the process of generating high frequencies by replicating low frequencies in the expanded spectrum introduced by upsampling. Since this creates duplicates in high-frequency bands, it's ability to control high frequencies is limited.

simply use data at a very high resolution, such that no high-frequency component remains, and train a larger scale CNN on the high resolution data. However, note that the larger scale CNN still contains the spatial frequency bias, which can be revealed when trained on a dataset with prominent high frequencies. For example, see Figure 1 where we train a large-scale StyleGAN2 (config-e) on a fractal dataset.

What all the aforementioned solutions have in common, is an appeal to the Universal Approximator Theorem: a neural network, equipped with hidden units and non-linearities, can model any continuous function given a large enough number of hidden units (*i.e*. increase in depth and/or width). However, in case of generative CNNs, even though the capacity of generating high frequencies can be increased to any desirable amount by the aforementioned solutions, low frequencies will always receive more capacity than high frequencies. This introduces a redundancy in generative CNNs, and comes at the cost of computational resources and generalization (a larger model demands more data). Naturally, one wonders if there is a way to more directly assign capacity to high frequencies, or to any spatial frequency of interest. The next section will explore this idea.

4 Frequency Shifted Generators (FSG)

In the previous section, we observed that GANs have a spatial frequency bias, favoring the learning of low frequencies, however, *is it possible to manipulate this bias such that it favors other frequencies?* If so, this would make it possible to explicitly target specific frequencies of interest in a dataset. In this section, we show how this can be achieved with minimal increase in training resources. Instead of inherently generating high frequencies, a generative model $G(x, y)$ can first generate a signal with prominent low frequencies and then transform the signal such that these prominent frequencies are shifted towards a desired frequency $(\hat{u}_t, \hat{v}_t)$. This can be achieved by a frequency shift operator:

$$G(x,y)e^{j2\pi(\hat{u}_t x+\hat{v}_t y)} = \frac{1}{mn}\sum_{u=0}^{m-1}\sum_{v=0}^{n-1} C(u,v)e^{j2\pi(\hat{u}+\hat{u}_t,\hat{v}+\hat{v}_t).(x,y)} \tag{5}$$

where $\hat{u}_t, \hat{v}_t \in [-0.5, 0.5)$. After the frequency shift, the frequency components previously close to $(0, 0)$ are now placed close to $(\hat{u}_t, \hat{v}_t)$. However, since G is generating a real signal and the spectrum of real signals are symmetric, it can not sufficiently represent a high-frequency band, that is, G can only represent symmetric frequency bands. Note that while natural images are real signals and have symmetric spectrum with respect to zero, a specific band of their spectrum is not necessarily symmetric. In order to generate a non-symmetric frequency band, we can use two neural networks to generate a real image (G_r) and an imaginary image (G_i), which together compose the complex generated image (G_c). The complex image is then shifted to $(\hat{u}_t, \hat{v}_t)$ according to Eq. (5) to construct the shifted generator G_s as follows:

$$G_s(x,y) = G_c(x,y)e^{j2\pi(\hat{u}_t x+\hat{v}_t y)} = \left[G_r(x,y) + jG_i(x,y)\right] e^{j2\pi(\hat{u}_t x+\hat{v}_t y)} \tag{6}$$

The real part of G_s is now generating an image which can sufficiently represent any frequency band, and has a spatial frequency bias favoring the desired component $(\hat{u}_t, \hat{v}_t)$:

$$\begin{aligned}\Re\left[G_s(x,y)\right] = {} & G_r(x,y)\cos(2\pi(\hat{u}_t x + \hat{v}_t y)) \\ & - G_i(x,y)\sin(2\pi(\hat{u}_t x + \hat{v}_t y))\end{aligned} \tag{7}$$

Frequency Shifted Generators (FSGs) can be used to efficiently target specific spatial frequency components in a dataset. Table 2 shows the results of training GANs using FSG with $(\hat{u}_t, \hat{v}_t) = (\frac{1}{2}, \frac{1}{2})$ on SCelebA and SBedrooms. The use of FSG has considerably improved the GANs' performance on these high-frequency datasets, with minimal increase in training resources. This also provides an interesting insight: the discriminator is able to effectively guide a capable generator towards learning high frequencies, therefore, the spatial frequency bias must be primarily rooted in the GAN's generator and not the discriminator. Moreover, multiple FSGs, with smaller network capacity, can be added to the main generator of GANs to improve performance on specific frequencies. Figure 4 shows the improvement in GANs trained on CelebA when their respective generators are enhanced by adding multiple FSGs with $(\hat{u}_t, \hat{v}_t)$ at $(\frac{1}{16}, 0)$, $(0, \frac{1}{16})$, $(-\frac{1}{16}, \frac{1}{16})$, and $(\frac{1}{16}, \frac{1}{16})$ (see Appendix for samples and details of the networks). Interestingly, the added FSGs specialize towards their respective target frequency $(\hat{u}_t, \hat{v}_t)$, without any explicit supervision during training. This provides further evidence of the spatial frequency bias: if unbiased, the added FSGs would have no incentive to specialize towards any specific frequency.

5 Related Works

Spectral Limitations of Neural Networks. Recent works on fully-connected neural networks have discovered a spectral bias against learning high-frequency functions (Rahaman et al. 2019; Basri et al. 2020), which can be addressed by using a proper high dimensional embedding of the input space (Tancik et al. 2020). However, while these works define a frequency component as a periodic change in a single output of the network with respect to changes in the input space, we define a frequency component as a periodic change across the adjacent outputs of the network (hence a *spatial* frequency component). Note that these two notions of frequency are independent by definition, that is, one can be mathematically defined while the other is not and vice versa, therefore a bias in one does not readily imply a bias in the other, and the arguments do not carry over.

Spectral Limitations of Generative CNNs. Spectral limitations have been observed in different tasks when using generative CNNs. In L2-reconstruction tasks, there is a bias against high frequencies, primarily attributed to the vanishing gradient of the L2 loss on low-power frequencies (Deng et al. 2020; Li and Barbastathis 2018; Ulyanov, Vedaldi, and Lempitsky 2018). In Auto Encoders (AEs) and Variational Auto Encoders (VAEs), there is a similar bias, primarily attributed to the distribution assumptions in their objectives (Huang et al. 2018; Larsen et al. 2016). In contrast, GAN's objective does not impose any such spectral limitations. In theory, GANs must be able to learn any suitable dis-

	SCelebA		SBedrooms	
Model	FID	LR (%)	FID	LR (%)
WGAN-GP	328.72 ± 9.70	59.04 ± 5.09	283.02 ± 7.06	42.42 ± 4.32
WGAN-FSG	20.70 ± 0.44	1.93 ± 0.57	59.81 ± 1.64	1.80 ± 0.28
PG-GAN	23.12 ± 2.08	3.93 ± 0.70	69.03 ± 10.28	3.12 ± 0.16
PG-GAN-FSG	17.91 ± 0.74	2.96 ± 0.55	54.64 ± 0.26	2.67 ± 0.75
StyleGAN2	343.57 ± 53.59	43.03 ± 34.10	260.84 ± 4.03	7.32 ± 1.86
StyleGAN2-FSG	7.17 ± 0.07	1.41 ± 0.10	67.85 ± 2.38	1.82 ± 0.27

Table 2: Performance gain (decrease in FID and LR) on the high-frequency datasets achieved by using FSG in GANs. Average with one standard deviation ($\pm$) is reported over three random training runs.

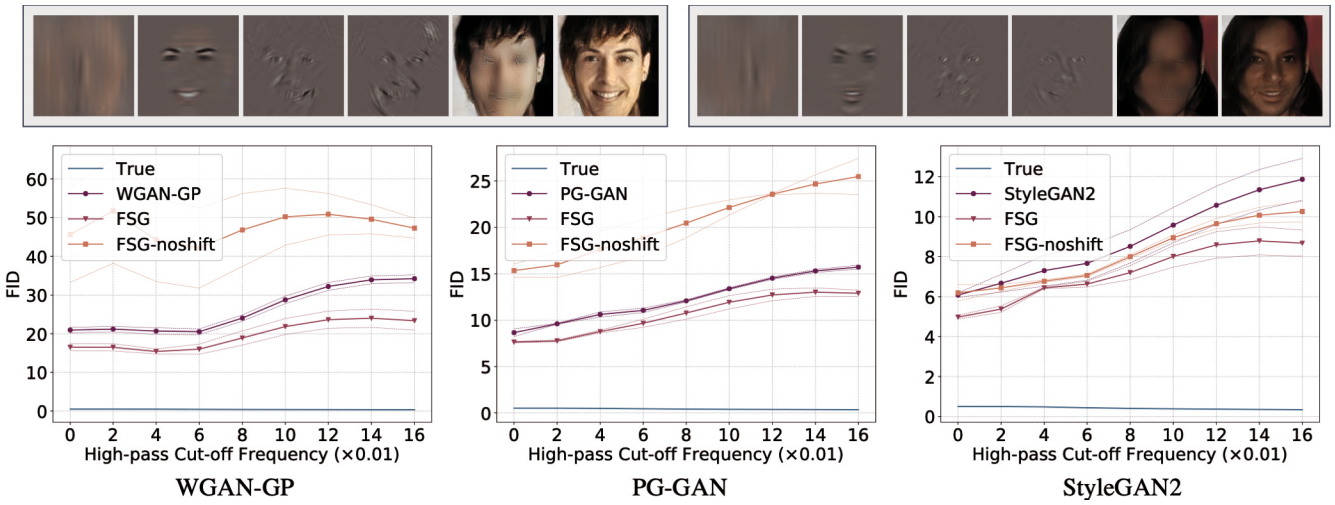

Figure 4: (Top) Two samples from WGAN-GP when enhanced by adding multiple FSGs, trained on CelebA. In each sample, from left to right, the outputs correspond to the FSG with $(\hat{u}_t, \hat{v}_t)$ at $(\frac{1}{16}, 0)$, $(0, \frac{1}{16})$, $(-\frac{1}{16}, \frac{1}{16})$, and $(\frac{1}{16}, \frac{1}{16})$, the WGAN-GP's main generator, and the final compound output (sum of all the preceding generators). Notice how each FSG has learned to focus on specific spatial frequencies, without any explicit supervision during training (see Appendix for more samples). (Bottom) Improvement in the FID Levels of multiple-FSG GANs trained on CelebA. Note that adding the same number of FSGs without the shift (FSG-noshift) does not yield the same improvements, showing the significance of the frequency shift.

tribution regardless of the carrier spatial frequencies. Therefore, while a spectral bias in generative CNNs could be obscured by the inherent spectral biases of AEs, VAEs and L2 tasks, GANs provide a clear lens for observing such biases.

Quantitative Metrics. The prevalent metrics for evaluating GANs, most notably Inception Score (Salimans et al. 2016), FID (Heusel et al. 2017), and MS-SSIM (Odena, Olah, and Shlens 2017), consider all spatial frequency components at once, thus lacking spectral resolution. Most relevant to our proposed metric, Karras et al. (2018) propose computing sliced Wasserstein distance between patches extracted from true and generated images at different levels of a Laplacian pyramid (SWD). Interestingly, evaluating GANs with SWD shows approximately similar performance across frequency bands (Karras et al. 2018). We conjecture that this discrepancy comes from the fact that small differences between patches in the pixel space, can result in large differences in the more meaningful feature space used by FID.

6 Conclusion

In this work, we showed the existence of a bias against high spatial frequencies in convolutional GANs, investigated its cause, and finally proposed a simple method illustrating that this bias is not fixed and can be effectively manipulated. Our findings suggest that the information carried by high frequencies is considerably more likely to be missed by GANs, a critical consideration when using GANs for data augmentation or reconstruction in applications concerned with intricate patterns, such as in medical and astronomy domains. We also observed that the spatial frequency bias primarily affects GAN's generator and not its discriminator. This gives the discriminator an advantage which can be the root of certain instabilities in GAN training. Investigating this connection between the spatial frequency bias and unstable GAN training, as well as extending Theorem 1 to incorporate the effect of various normalization and stabilization techniques, are interesting directions for future research.

References

Arjovsky, M.; and Bottou, L. 2017. Towards principled methods for training generative adversarial networks. *arXiv preprint arXiv:1701.04862*.

Arora, S.; Ge, R.; Liang, Y.; Ma, T.; and Zhang, Y. 2017. Generalization and equilibrium in generative adversarial nets (gans). In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 224–232. JMLR. org.

Basri, R.; Galun, M.; Geifman, A.; Jacobs, D.; Kasten, Y.; and Kritchman, S. 2020. Frequency bias in neural networks for input of non-uniform density. In *International Conference on Machine Learning*, 685–694. PMLR.

Deng, M.; Li, S.; Goy, A.; Kang, I.; and Barbastathis, G. 2020. Learning to synthesize: robust phase retrieval at low photon counts. *Light: Science & Applications*, 9(1): 1–16.

Durall, R.; Keuper, M.; and Keuper, J. 2020. 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*, 7890–7899.

Dzanic, T.; Shah, K.; and Witherden, F. 2020. Fourier Spectrum Discrepancies in Deep Network Generated Images. In *Advances in Neural Information Processing Systems*.

Field, D. J. 1987. Relations between the statistics of natural images and the response properties of cortical cells. *Josa a*, 4(12): 2379–2394.

Gibbs, A. L.; and Su, F. E. 2002. On choosing and bounding probability metrics. *International statistical review*, 70(3): 419–435.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, 2672–2680.

Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; and Courville, A. C. 2017. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, 5769–5779.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, 6626–6637.

Huang, H.; He, R.; Sun, Z.; Tan, T.; et al. 2018. Introvae: Introspective variational autoencoders for photographic image synthesis. In *Advances in neural information processing systems*, 52–63.

Huang, X.; Li, Y.; Poursaeed, O.; Hopcroft, J.; and Belongie, S. 2017. Stacked generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5077–5086.

Karras, T.; Aila, T.; Laine, S.; and Lehtinen, J. 2018. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *International Conference on Learning Representations*.

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

Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2020. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8110–8119.

Khayatkhoei, M.; Singh, M. K.; and Elgammal, A. 2018. Disconnected manifold learning for generative adversarial networks. In *Advances in Neural Information Processing Systems*, 7343–7353.

Larsen, A. B. L.; Sønderby, S. K.; Larochelle, H.; and Winther, O. 2016. Autoencoding beyond pixels using a learned similarity metric. In *International conference on machine learning*, 1558–1566. PMLR.

Li, S.; and Barbastathis, G. 2018. Spectral pre-modulation of training examples enhances the spatial resolution of the phase extraction neural network (PhENN). *Optics express*, 26(22): 29340–29352.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Odena, A.; Olah, C.; and Shlens, J. 2017. Conditional image synthesis with auxiliary classifier gans. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 2642–2651. JMLR. org.

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

Rahaman, N.; Baratin, A.; Arpit, D.; Draxler, F.; Lin, M.; Hamprecht, F.; Bengio, Y.; and Courville, A. 2019. On the spectral bias of neural networks. In *International Conference on Machine Learning*, 5301–5310. PMLR.

Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; and Chen, X. 2016. Improved techniques for training gans. In *Advances in neural information processing systems*, 2234–2242.

Tancik, M.; Srinivasan, P. P.; Mildenhall, B.; Fridovich-Keil, S.; Raghavan, N.; Singhal, U.; Ramamoorthi, R.; Barron, J. T.; and Ng, R. 2020. Fourier Features Let Networks Learn High Frequency Functions in Low Dimensional Domains. *NeurIPS*.

Ulyanov, D.; Vedaldi, A.; and Lempitsky, V. 2018. Deep image prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9446–9454.

Wang, T.-C.; Liu, M.-Y.; Zhu, J.-Y.; Tao, A.; Kautz, J.; and Catanzaro, B. 2018. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8798–8807.

Yu, F.; Zhang, Y.; Song, S.; Seff, A.; and Xiao, J. 2015. LSUN: Construction of a Large-scale Image Dataset using Deep Learning with Humans in the Loop. *arXiv preprint arXiv:1506.03365*.