SIMPLER IS BETTER: SPECTRAL REGULARIZATION AND UP-SAMPLING TECHNIQUES FOR VARIATIONAL AUTOENCODERS

Sara Björk[1], *Jonas Nordhaug Myhre*[2,1] *and Thomas Haugland Johansen*[2,1]

[1]UiT The Arctic University of Norway, [2]NORCE Norwegian Research Centre

ABSTRACT

Full characterization of the spectral behavior of generative models based on neural networks remains an open issue. Recent research has focused heavily on generative adversarial networks and the high-frequency discrepancies between real and generated images. The current solution to avoid this is to either replace transposed convolutions with bilinear up-sampling or add a spectral regularization term in the generator. We propose a 2D Fourier transform-based spectral regularization loss and evaluate it on the variational autoencoder. We show that it can achieve results equal to, or better than, the current state-of-the-art in frequency-aware losses for generative models. In addition, we experiment with altering the up-sampling procedure in the generator network and investigate how it influences the spectral performance of the model. We include experiments on synthetic and real data sets to demonstrate our results.

1. INTRODUCTION

Generative deep neural network models such as the Generative Adversarial Network (GAN) [1] and the Variational Autoencoder (VAE) [2] have in recent years gained a lot of attention in e.g. face generation [3, 4, 5], image-to-image translation or style-transfer [6, 7, 8] tasks. The wide applicability of generative models has fostered a large body of research that aims to improve generative network architectures to enhance the quality of the generated images. Most of this work has focused on variations of spatial loss terms in the objective functions, which has led to a multitude of different GAN and VAE architectures, see e.g. [9, 10, 11, 8, 12, 13, 14, 15]. Although current methods generate very realistic-looking natural images, see e.g. [3, 4, 5], generative neural network models are in general not able to reproduce the spectral distribution of natural images adequately. Generated images still suffer from blurriness and lack of sharp details. Figure. 1 illustrates this issue in columns (a) and (b), where (a) shows an original sample from the CelebA dataset [16], and (b) is a blurry reconstruction of the same sample image from a VAE trained with a traditional spatial objective function (Vanilla VAE).

The lack of high-frequency content can be partially explained by the *spectral bias* of neural networks [17]; neural networks prioritize low-frequency components of the data in the early stages of learning. A growing body of research has investigated these findings, see e.g. [18, 19, 20, 21], and ways to utilize this in e.g. deep-fake detection [22, 23, 24, 25]. Others propose different ways to resolve, work around, or reduce the effects of a bias towards the low-frequency components [23, 26, 24, 27, 26, 19, 28, 29, 30]. Another partial explanation for the discrepancy in the frequency content of generated images is the transposed convolution operation used in the up-sampling components of generative models. Durall et al. [23] argue that the transposed convolution operation is causing the models' inability to learn the high-frequency content of the data and propose to append *spectral regularization* (SR) to the spatial objective function to mitigate the effects caused by the up-sampling strategy. Others, see e.g. [19, 28], suggest to replace the last up-sampling operation in the architecture. In this work we show that a simple frequency-aware loss that forces the generative model to focus on agreement of the overall spectral content of the data is equally effective, and sometimes better than the current state-of-the-art-in SR [23, 26]. Furthermore, we consider the effects of replacing the up-sampling operation in the last layer, similar to [19, 28]. This allows us to

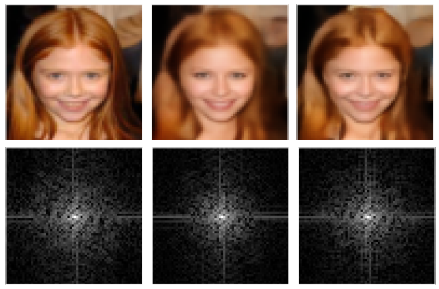


Fig. 1: A sample from the CelebA dataset [16]. Top row: **(a)**: real image, **(b)**: Vanilla VAE reconstruction and **(c)**: reconstruction from VAE with spectral regularization. Bottom row: FFT spectrum of the corresponding images. We note the discrepancies at the highest frequencies of the 2D Fourier spectrum in (b), compared to (a), and the lack of details in the spatial representation of the image. A simple 2D FFT regularization (c) achieves less blurriness in the spatial domain and less discrepancies in the Fourier spectrum.

evaluate the impact of the up-sampling operation both with and without SR. Concretely, we propose to incorporate a simple 2D Fourier transform agreement loss term, denoted the FFT loss, into the overall objective function. With this additional spectral agreement term, we wish to align the high-frequency components of the Fourier spectrum while penalizing unilateral learning of low-frequency components of the data. We limit our focus to VAE models as they generally are easier to train than GANs. Therefore, they can serve as a simpler baseline, used to investigate the spectral properties of neural network-based generative models. However, as the FFT loss complements existing spatial losses, its utility is not restricted to VAEs only. Figure. 1 shows that inclusion of SR through the proposed loss results in better spatial and spectral agreement with the original data, i.e. compare the Vanilla VAE in column (b) with the proposed loss in (c). We empirically evaluate the FFT loss on synthetic and real datasets, and compare it with two more complex SR methods; the azimuthal integration loss by Durall et al. [23] and the Watson perceptual loss from Czolbe et al. [26]. All three SR methods are compared against the baseline VAE objective function with its binary cross-entropy (BCE) loss.

The rest of the paper is organized as follows: we review some related work that focuses on spectral reconstruction with generative models in Sec. 1.1. Sec. 2 briefly introduce the VAE, SR with our FFT agreement loss, and introduce an alternative approach to the transposed convolution up-sampling operation. Results from our experiments are presented and discussed in Sec. 3. Sec. 4 concludes this work with a summary of our most important findings.

1.1. Related Work

Many works have illustrated the problems of generative models and spectral reconstruction. Several theories exist, but the most notable are *spectral bias* [17, 18] and issues related to the up-sampling operations in the final layers of the generator network [23, 28]. Karras et al. [4] generate high-resolution images by first letting their network focus on low-resolution images and then progressively shift the training to consider higher-resolution images. However, as pointed out by Khayatkhoei and Elgammal [18], application of the StyleGAN2 [4], which samples at high frequencies, might avoid the spatial frequency bias without actually solving the issue: high-frequency components, such as sharp details, are not preserved to the same extent in data that has been sampled at very high resolution [18]. Access to high-definition or high-resolution images is not always possible, especially not when working with e.g. remote sensing data or medical data. Very deep architectures might also be unsuitable when considering available computational power or computation time in specific applications.

There are numerous works in the last two years that either try to explain the frequency discrepancy from a theoretical perspective, e.g. [31, 20], or acknowledge this drawback by proposing ways to resolve or reduce the effects of the spectral bias. Particularly important is the work by Durall et al. [23], which illustrates how standard up-sampling methods such as up-convolution or transposed convolution in the generator network result in generative models that are incapable of reproducing the spectral distribution of the data. While up-sampling methods lack high frequencies, transposed convolution, on the other hand, adds a large amount of high-frequency noise in the up-sampling process. To overcome the issues the up-sampling process causes, they propose to include SR in the generator's space-based loss. The spectral part of the loss is a 1D representation of the Fourier power spectrum given by azimuthal integration over the radial frequencies. Czolbe et al. [26] also adopted the idea of SR by proposing a loss function based on Watson's visual perception model [32]. It mimics the human perception of image data using a weighted distance in frequency space and adjusting for contrast and luminance. Jiang et al. [29] propose a dynamic loss function that focuses on frequencies that are hard to learn by, during training, let a spectrum weight matrix down-weight easy frequencies[1]. From now on, we refer to SR by azimuthal integration [23] as the AI loss, and to [26] as the Watson-DFT loss. In [19], Chandrasegaran et al. argue that the spectral discrepancies are not inherent to the neural network, but an artifact from the up-sampling procedure. They show promising results by replacing the last transposed convolution layer with either zero-insert scaling, nearest interpolation, or bilinear interpolation followed by traditional convolution.

2. METHODOLOGY

This section briefly introduces VAEs and the proposed frequency-aware loss. A section on common up-sampling procedures in convolutional neural networks is also included.

2.1. Variational Autoencoders

A VAE is a Bayesian generative model configured in an autoencoding fashion, with an encoder mapping the data, x, into a latent variable, z, and a decoder that maps the latent variable back to the original data space. As usual in a Bayesian setup, the problem of inference is to find the posterior distribution $p(z|x)$. Since the evidence $p(x)$ is typically intractable, a lower bound is optimized using variational inference [2]:

$$\arg\min_{\phi,\theta} \mathbb{E}_{q_\phi(z|x)} \left\{\log p_\theta(x|z)\right\} - \beta \mathrm{KL}\left(q_\phi(z|x)||p(z)\right). \quad (1)$$

For full derivations, see [33] or [34]. Both $q_\phi(z|x)$ (encoder) and $p_\theta(x|z)$ (decoder) are modeled via neural networks. $p(z)$ is the prior over the latent variable z, which is commonly assumed to be multivariate Gaussian distributed. By identifying

[1] Although the work by Jiang et al. [29] is similar to our work, their loss is more complex as their spectrum weight matrix has to be computed for every single image. Comparing our FFT loss to [29] is deferred to future work due to space limitations.

$\log p_\theta(x|z)$ as the negative BCE loss, we can replace it with an energy-based model, $p(x|z) \propto \exp\{-L(x, \mu_x(z))\}$ where L is any function that leads to a proper probability density function [26]. This formulation allows alternative reconstruction losses, such as the Watson perceptual loss used in [26].

2.2. A 2D frequency spectral regularization loss

Both Czolbe et al. [26] and Durall et al. [23] suggest that the combination of spectral and spatial components in the reconstruction loss helps to improve the image quality of generated samples. Motivated by them, we propose a simple SR, the FFT loss function[2]:

$$\begin{aligned}\mathcal{L}_f(x, \hat{x}) = &\frac{1}{n} \sum \left(imag[\mathcal{F}\{x\}] - imag[\mathcal{F}\{\hat{x}\}]\right)^2 + \\ &\frac{1}{n} \sum \left(real[\mathcal{F}\{x\}] - real[\mathcal{F}\{\hat{x}\}]\right)^2,\end{aligned} \tag{2}$$

that through the MSE enforce alignment of the *real* and *imaginary* parts of the actual image, *x*, and its reconstruction, $\hat{x}$, in the frequency domain. $\mathcal{F}$ denotes the fast Fourier Transform and n is the total number of pixels in the image. The proposed FFT loss is complementary to existing spatial losses and can be used by generative models through:

$$\mathcal{L}(x, \hat{x}) = (1 - \alpha)\mathcal{L}_s(x, \hat{x}) + \alpha\mathcal{L}_f(x, \hat{x}), \quad \alpha \in [0, 1]. \tag{3}$$

where $\mathcal{L}_s(x, \hat{x})$ is e.g. the BCE loss, computed in the spatial domain between *x* and $\hat{x}$ and α determines the relative influence of $\mathcal{L}_s$ and $\mathcal{L}_f$. We apply our FFT loss for the VAE by replacing L in Sec. 2.1 with Eq. 3.

2.3. Up-sampling and transposed convolution

To convert the latent vector into a higher-dimensional output space, e.g., transforming a low-dimensional Gaussian sample to an image, the generator needs to increase the resolution in each layer. The most common strategy is to use the *transposed convolution* operation, where the input is zero-padded and convolved with the appropriate filter. See [35] for a complete description. The alternative approach is to split the transformation in two: *up-sampling by interpolation* and *convolution*. The transposed convolution operation is known to have several shortcomings, such as high-frequency discrepancies and checkerboard artifacts [19, 36]. Chandrasegaran et al. [19] propose multiple ways to perform the up-sampling in the last layer of a generator network. Their results advise to use up-sampling with nearest-neighbor interpolation and a single convolutional block of kernel size 5×5. We adopt this setup in this work, hereafter denoted 'N.1.5', and refer to [19] for a comprehensive evaluation of additional versions of the up-sampling procedure.

[2]We also evaluated a FFT loss similar to [30], that used the *amplitude* and *phase* information, but this did not perform as good as the proposed loss in Eq. 2.

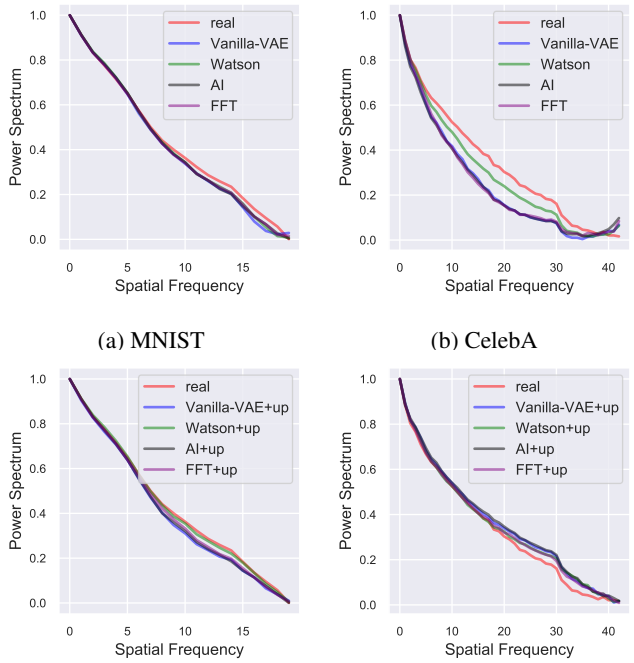
3. EXPERIMENTS

We evaluate SR through our proposed 2D spectral loss and compare its performance to the AI loss [23], the Watson-DFT loss [26], and VAE without SR (Vanilla-VAE). Both our proposed FFT loss and the Watson-DFT loss can handle gray-scale and RGB images without any modification of the data. However, the original AI loss can only handle gray-scale images, implying that RGB images first need to be transformed into gray-scale images [23]. We evaluated the possible impact of their gray-scale transformation but found it to be neglectable.

We employ three different datasets with increasing complexity: a simple gray-scale version of the Shape dataset by Jing et al. [37], MNIST (grey-scale) [38], and CelebA (RGB) [16]. The Shape and MNIST datasets were analyzed at 32x32 resolution, while the CelebA dataset was analyzed at 64x64 resolution. We chose to employ the VAE networks from [37]. The focus of this work is to evaluate how SR, either alone or combined with last layer up-sampling [19], can enhance image quality. The influence of different network architectures on the models' ability to reproduce the high-frequency content of the data is beyond the scope of this work and has therefore been omitted. The interested reader could consult [3, 4, 5] for examples of generative model architectures focused on generating high-resolution images from low-resolution images. We evaluate the performance of each model by the root mean squared error (RMSE) and azimuthal (polar coordinate) integration of the Fourier spectrum. Since the AI loss focuses on alignment in the 1D representation of the Fourier power spectrum, the Vanilla-VAE on the alignment in the spatial domain, and our FFT loss on the alignment in the 2D representation of the Fourier spectrum, we choose to compute the RMSE in all these three domains. RMSE metrics for the Watson-DFT loss on the Shape dataset have been omitted from the reported results of Sec. 3.1 and Sec. 3.2, since these models did not work correctly. This is most likely because the Shape dataset is too simple for a perceptual loss function.

3.1. Spectral regularization with transposed convolution

We trained VAE models for each of the three different datasets with the baseline Vanilla-VAE alone, or together with the Watson-DFT, AI or the FFT loss, by employing traditional transposed convolution up-sampling. Our purpose was to compare the three different ways to achieve SR against each other and the baseline. Based on this, we evaluated if our contribution, the FFT loss, can compete against the Watson-DFT and the AI loss. Table. 1 summarizes the quantitative results from the empirical investigation, while Figure. 2a and Figure. 2b show the average azimuthal integration of the power spectrum for models trained with the different objective functions. The quantitative RMSE metrics in Table. 1 show that our proposed FFT loss performs well in generating images that resemble the true images. The only exception is when the RMSE is



(a) MNIST (b) CelebA

(c) MNIST+up (d) CelebA+up

Fig. 2: Average azimuthal integration power spectrum computed for images in a test batch of either the MNIST [38] (first column) or the CelebA dataset [16] (second column) by applying either the Vanilla-VAE alone, or together with SR. Models trained with the traditional transposed convolution up-sampling operation are shown in **(a)** and **(b)**. Models trained with the 'N.1.5' up-sampling [19] are shown in **(c)** and **(d)**.

computed in the AI-domain, where the Watson-DFT loss has a smaller RMSE than our FFT loss for the CelebA dataset. This is also shown in Figure. 2b.

3.2. Different last-layer up-sampling procedures

We changed the up-sampling of the last layer of the VAEs from the traditional transposed convolution to 'N.1.5', as introduced in Sec. 2.3, and repeated the experiments from Sec. 3.1. Table. 2 summarizes the quantitative results from the empirical investigation, while Figure. 2c and Figure. 2d show average azimuthal integration of the power spectrum for models trained on either the MNIST or the CelebA dataset. Results in Table. 2 show that models trained with our proposed FFT loss resemble the true data distribution better than any of the other evaluated objective functions. Comparing the rightmost parts of Figure. 2a and Figure. 2b to Figure. 2c and Figure. 2d, it can be noted that the change in up-sampling to 'N.1.5' improves the alignment of the high frequencies for all generative models. However, it should be noted that the change in up-sampling does not always imply lower RMSE, compare e.g. Watson-DFT for CelebA in the AI-domain in Table. 1 with the

Table 1: Mean ± std RMSE computed in AI-domain, 2D Fourier Transform (FT) domain and spatial domain, for experiments in Sec. 3.1 with transposed convolution up-sampling.

Objective function	Dataset	AI	2D FT	Spatial domain
Vanilla-VAE:	Shape	1.2226 ± 0.8317	1.1791 ± 0.3184	0.0004 ± 0.0088
AI:	Shape	0.8892 ± 0.4943	1.0842 ± 0.3549	0.0003 ± 0.0055
FFT:	Shape	**0.4447 ± 0.1671**	**1.0543 ± 0.3289**	**0.0001 ± 0.0021**
Vanilla-VAE:	MNIST	3.5091 ± 0.8296	1.7728 ± 0.1371	0.0081 ± 0.1331
Watson-DFT:	MNIST	3.9728 ± 0.9697	1.8497 ± 0.0962	0.0094 ± 0.0282
AI:	MNIST	3.5084 ± 0.9108	1.7274 ± 0.1194	0.0079 ± 0.0228
FFT:	MNIST	**2.8761 ± 0.5774**	**1.5710 ± 0.1249**	**0.0062 ± 0.0174**
Vanilla-VAE:	CelebA	9.2300± 1.8732	4.3506 ± 0.7480	0.0323± 0.0376
Watson-DFT:	CelebA	**6.7433 ± 2.7038**	4.0370 ± 0.7803	0.0284 ± 0.0356
AI:	CelebA	9.5100 ± 2.0860	4.3345 ± 0.6696	0.0316 ± 0.0372
FFT:	CelebA	8.3299 ± 1.5421	**3.5406 ± 0.6743**	**0.0237 ± 0.0303**

Table 2: Mean ± std RMSE for experiments in Sec. 3.2 with 'N.1.5' up-sampling in the last layer, following [19].

Objective function	Dataset	AI	2D FT	Spatial domain
Vanilla-VAE:	Shape	1.3893 ± 0.7328	1.1983 ± 0.3240	0.0004 ± 0.0083
AI:	Shape	0.8041 ± 0.4916	1.0940 ± 0.3186	0.0002 ±0.0046
FFT:	Shape	**0.1394 ± 0.1103**	**1.0325 ± 0.3248**	**2.6892e-05 ± 0.0005**
Vanilla-VAE:	MNIST	3.4711 ± 1.0887	1.8513 ± 0.1008	0.0084 ±0.0234
Watson-DFT:	MNIST	3.6987 ± 1.2195	1.9171 ±0.1164	0.0100 ± 0.0278
AI:	MNIST	3.1265 ±0.9568	1.8270 ± 0.0848	0.0085 ± 0.0236
FFT:	MNIST	**2.9276 ± 0.9551**	**1.7041 ± 0.0958**	**0.0071 ± 0.0189**
Vanilla-VAE:	CelebA	9.4452 ± 3.7583	4.3022 ± 0.7461	0.0315 ± 0.0370
Watson-DFT:	CelebA	6.8082 ±2.5990	4.0807 ± 0.8466	0.0290± 0.0359
AI:	CelebA	8.5200 ±4.0584	3.6181 ± 4.3116	0.7356 ± 0.0367
FFT:	CelebA	**5.8252 ± 2.5366**	**3.5917 ± 0.6925**	**0.0245 ± 0.0315**

corresponding value in Table. 2. For the AI loss and CelebA we can verify the results from [19]; changing the up-sampling to 'N.1.5' reduces the RMSE in both the AI-domain and the 2D Fourier transform-domain. However, this result is not consistent for all datasets, over all tested SR losses, nor for the baseline Vanilla-VAE. This indicates that a change in the up-sampling procedure is one way to improve the performance of generative models, but the effect is not consistent, and we urge more research on this topic.

4. CONCLUSION AND FUTURE WORK

In this paper, we have shown that a simple spectral regularization term based on the 2D Fourier transform performs better than more complex regularization methods for improving the image quality of the VAE generative model. Moreover, our results show that changing the up-sampling procedure in the last layer from transposed convolution to nearest-neighbor interpolation followed by standard convolution gives more ambiguous results than indicated by previous research. Clearly, more research is needed to untangle the true spectral properties of neural generative models.

5. ACKNOWLEDGEMENTS

We thank Stian Normann Anfinsen at NORCE and Robert Jenssen at UiT Machine Learning Group for their valuable feedback.

6. REFERENCES

[1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative Adversarial Nets," in *NeurIPS*, 2014, pp. 2672–2680.

[2] Diederik P Kingma and Max Welling, "Auto-Encoding Variational Bayes," in *Proc. of ICLR*, 2014.

[3] Tero Karras, Samuli Laine, and Timo Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks," *Proc. of IEEE CVPR*, vol. 2019-June, pp. 4396–4405, 2019.

[4] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila, "Analyzing and Improving the Image Quality of StyleGAN," *Proc. of IEEE CVPR*, pp. 8107–8116, 2020.

[5] Arash Vahdat and Jan Kautz, "NVAE: A Deep Hierarchical Variational Autoencoder," in *NeurIPS*, 2020, p. 19667–19679.

[6] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros, "Image-To-Image Translation With Conditional Adversarial Networks," in *Proc. of IEEE CVPR*, 2017, pp. 1125–1134.

[7] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha, "StarGAN v2: Diverse Image Synthesis for Multiple Domains," in *Proc. of IEEE CVPR*, 2020, pp. 8188–8197.

[8] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros, "Unpaired Image-To-Image Translation Using Cycle-Consistent Adversarial Networks," in *Proc. of IEEE ICCV*, 2017, pp. 2223–2232.

[9] Alec Radford, Luke Metz, and Soumith Chintala, "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks," in *Proc. of ICLR*, 2015.

[10] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville, "Improved Training of Wasserstein GANs," *NeurIPS*, vol. 30, pp. 5767–5777, 2017.

[11] Xudong Mao, Qing Li, Haoran Xie, Raymond Y K Lau, Zhen Wang, and Stephen Paul Smolley, "Least Squares Generative Adversarial Networks," in *Proc. of IEEE ICCV*, 2017, pp. 2794–2802.

[12] Rewon Child, "Very Deep VAEs Generalize Autoregressive Models and Can Outperform Them on Images," in *Proc. of ICLR*, 2021.

[13] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther, "Autoencoding beyond pixels using a learned similarity metric," in *Proc. of ICML*. 2016, pp. 1558–1566, PMLR.

[14] Aaron Van Den Oord, Oriol Vinyals, and Koray Kavukcuoglu, "Neural Discrete Representation Learning," *NeurIPS*, vol. 2017-Decem, no. Nips, pp. 6307–6316, 2017.

[15] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals, "Generating Diverse High-Fidelity Images with VQ-VAE-2," *NeurIPS*, vol. 32, 2019.

[16] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang, "Deep Learning Face Attributes in the Wild," in *Proc. of IEEE ICCV*, 2015, pp. 3730–3738.

[17] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville, "On the Spectral Bias of Neural Networks," in *Proc. of ICML*. 2019, pp. 5301–5310, PMLR.

[18] Mahyar Khayatkhoei and Ahmed Elgammal, "Spatial Frequency Bias in Convolutional Generative Adversarial Networks," *arXiv preprint arXiv:2010.01473*, 2020.

[19] Keshigeyan Chandrasegaran, Ngoc-Trung Tran, and Ngai-Man Cheung, "A Closer Look at Fourier Spectrum Discrepancies for CNN-generated Images Detection," in *Proc. of IEEE CVPR*, 2021, pp. 7200–7209.

[20] Haohan Wang, Xindi Wu, Zeyi Huang, and Eric P Xing, "High-Frequency Component Helps Explain the Generalization of Convolutional Neural Networks," in *Proc. of IEEE CVPR*, 2020, pp. 8684–8694.

[21] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein, "Implicit Neural Representations with Periodic Activation Functions," *NeurIPS*, vol. 33, 2020.

[22] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz, "Leveraging Frequency Analysis for Deep Fake Image Recognition," *Proc. of ICML*, vol. PartF16814, pp. 3205–3216, 2020.

[23] Ricard Durall, Margret Keuper, and Janis Keuper, "Watch Your Up-Convolution: CNN Based Generative Deep Neural Networks are Failing to Reproduce Spectral Distributions," in *Proc. of IEEE CVPR*, 2020, pp. 7890–7899.

[24] Tarik Dzanic, Karan Shah, and Freddie Witherden, "Fourier Spectrum Discrepancies in Deep Network Generated Images," *NeurIPS*, vol. 33, 2020.

[25] Steffen Jung and Margret Keuper, "Spectral Distribution Aware Image Generation," in *Proc. of AAAI*, 2021, vol. 35, pp. 1734–1742.

[26] Steffen Czolbe, Oswin Krause, Ingemar Cox, and Christian Igel, "A Loss Function for Generative Neural Networks Based on Watson's Perceptual Model," *NeurIPS*, vol. 33, 2020.

[27] Yuanqi Chen, Ge Li, Cece Jin, Shan Liu, and Thomas H Li, "SSD-GAN: Measuring the Realness in the Spatial and Spectral Domains," in *Proc. of AAAI*, 2021, pp. 1105–1112.

[28] Yuehui Wang, Liyan Cai, Dongyu Zhang, and Sibo Huang, "The Frequency Discrepancy Between Real and Generated Images," *IEEE Access*, vol. 9, pp. 115205–115216, 2021.

[29] Liming Jiang, Bo Dai, Wayne Wu, and Chen Change Loy, "Focal Frequency Loss for Image Reconstruction and Synthesis," in *Proc. of IEEE CVPR*, 2021, pp. 13919–13929.

[30] Dario Fuoli, Luc Van Gool, and Radu Timofte, "Fourier space losses for efficient perceptual image super-resolution," in *Proc. of IEEE ICCV*, 2021, pp. 2360–2369.

[31] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng, "Fourier Features Let Networks Learn High Frequency Functions in Low Dimensional Domains," *NeurIPS*, vol. 33, 2020.

[32] Andrew B Watson, "DCT Quantization Matrices Visually Optimized for Individual Images," in *Human Vision, Visual Processing, and Digital Display IV*. 1993, vol. 1913, pp. 202–216, International Society for Optics and Photonics.

[33] David M Blei, Alp Kucukelbir, and Jon D McAuliffe, "Variational Inference: A Review for Statisticians," *Journal of the American statistical Association*, vol. 112, no. 518, pp. 859–877, 2017.

[34] Diederik P Kingma and Max Welling, "An Introduction to Variational Autoencoders," *arXiv preprint arXiv:1906.02691*, 2019.

[35] Vincent Dumoulin and Francesco Visin, "A guide to convolution arithmetic for deep learning," *arXiv preprint arXiv:1603.07285*, 2016.

[36] Augustus Odena, Vincent Dumoulin, and Chris Olah, "Deconvolution and Checkerboard Artifacts," *Distill*, vol. 1, no. 10, pp. e3, 2016.

[37] Li Jing, Jure Zbontar, and Yann LeCun, "Implicit Rank-Minimizing Autoencoder," *NeurIPS*, vol. 33, 2020.

[38] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner, "Gradient-Based Learning Applied to Document Recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.