Spectral Distribution Aware Image Generation

Steffen Jung, [1] **Margret Keuper** [2]

[1] Max Planck Institute for Informatics, Saarland Informatics Campus
[2] Data and Web Science Group, University of Mannheim
steffen.jung@mpi-inf.mpg.de, keuper@uni-mannheim.de

Abstract

Recent advances in deep generative models for photo-realistic images have led to high quality visual results. Such models learn to generate data from a given training distribution such that generated images can not be easily distinguished from real images by the human eye. Yet, recent work on the detection of such fake images pointed out that they are actually easily distinguishable by artifacts in their frequency spectra. In this paper, we propose to generate images according to the frequency distribution of the real data by employing a spectral discriminator. The proposed discriminator is lightweight, modular and works stably with different commonly used GAN losses. We show that the resulting models can better generate images with realistic frequency spectra, which are thus harder to detect by this cue.

Introduction

Image generation using generative adversarial networks has made huge progress in recent years. Especially the generation of photo-realistic images at high resolution has arrived at a level where it becomes hard for humans to distinguish between real and generated images. While the training data distribution appears to be well learned in the model's latent space, it is surprising how reliably real and generated images can be distinguished even when various cloaking techniques such as blurring or compression are applied (Yu, Davis, and Fritz 2019). In recent works on the detection of generated images (Frank et al. 2020; Durall, Keuper, and Keuper 2020), it was argued that this effect, at least partially, is due to artifacts introduced during the generation process itself. Commonly used up-sampling schemes in fact seem to generate these artifacts which are mostly in the high frequency domain and can not be corrected by the network itself (Frank et al. 2020; Durall, Keuper, and Keuper 2020; Bai et al. 2020). Such artifacts are undesired, not only because generated images can easily be identified as such, but also because they might be perceivable for example as grid patterns on the generated images.

Durall, Keuper, and Keuper (2020) propose a GAN regularization approach as a remedy. They argue that a generator network, given a sufficient amount of convolutional layers, can generate images with realistic frequency spectra if

they are penalized for deviating from the average frequency spectrum of the real images during training. In fact, the proposed regularization not only allowed to produce more realistic frequency spectra but also had a stabilizing effect on the training. However, the resulting spectra were still not able to match the distribution of the real data.

In this paper, we address this problem in a different way. Instead of introducing a regularization term, we propose to use a second discriminator which directly acts on the power spectra of real and generated images. This way, the generator network is not forced to produce images with average power spectra as in (Durall, Keuper, and Keuper 2020) but is enabled to learn the distribution of frequency spectra from training data. Since the training of GANs is computationally expensive we argue that an additional discriminator should be lightweight so that diverse architectures can easily adopt it. We therefore base our discriminator on one-dimensional (1D) projections of the frequency spectra instead of acting on the full two-dimensional data. We show that the resulting model can be trained with different commonly used GAN losses and evaluate its ability to fit the real images' frequency spectra in terms of a proposed cloaking score as well as in terms of the performance of frequency based generated image detection (Durall, Keuper, and Keuper 2020).

Related Work

Generative convolutional neural networks have recently been successful in a wide range of applications, such as the generation of photo-realistic images at high resolution (Karras et al. 2017; Brock, Donahue, and Simonyan 2018a; Karras, Laine, and Aila 2019a; Karras et al. 2019) to style transfer (Isola et al. 2017; Zhu et al. 2017a,b; Huang et al. 2018), or more generally image-to-image translation (Pathak et al. 2016; Iizuka, Simo-Serra, and Ishikawa 2017; Zhu et al. 2017b; Choi et al. 2018; Mo, Cho, and Shin 2019; Karras, Laine, and Aila 2019b) and text-to-image translation (Reed et al. 2016; Dai et al. 2017; Zhang et al. 2017, 2018). Generative adversarial networks (GANs) (Goodfellow et al. 2014) play a crucial role in this context. They aim to approximate a latent-space model of the underlying data distributions from training images in a zero-sum game between a generator and a discriminator network. From this latent data distribution model, new samples can be generated (drawn) by sampling. Recent works towards improving GANs pro-

posed different loss functions, regularizations or latent space constraints (Gulrajani et al. 2017a; Mao et al. 2017; Gulrajani et al. 2017b; Miyato et al. 2018; Mirza and Osindero 2014; Donahue, Krähenbühl, and Darrell 2016; Gurumurthy, Kiran Sarvadevabhatla, and Venkatesh Babu 2017; Durall, Keuper, and Keuper 2020; Brock, Donahue, and Simonyan 2018b; Kodali et al. 2017) to improve training stability and aim at high image resolutions (Karras et al. 2017; Karras, Laine, and Aila 2019a; Karras et al. 2019).

Since, with this progress, generated images are hard to distinguish from real images by the human eye, recent work has gained interest in the detection of generated images. On the one end, automatically detecting generated images helps to protect content authenticity in the context of deep fakes. On the other hand, it can help to improve the generation process itself as it allows to find systematic mistakes currently made by image generation networks. One such systematic mistake seems to be especially apparent in the frequency spectra of generated images (Durall et al. 2019; Wang et al. 2020; Frank et al. 2020; Bai et al. 2020). By feeding the Fourier transform or the discrete cosine transform of generated images into a deep network (Zhang, Karaman, and Chang 2019) or simpler learning models such as support vector machines (Durall et al. 2019) or ridge regression (Frank et al. 2020), surprisingly high detection rates can be achieved. As analyzed for example in (Durall et al. 2019; Frank et al. 2020), these systematic artifacts in the frequency domain are an effect of the generation process itself, more precisely in the up-convolutions. In (Durall, Keuper, and Keuper 2020), a regularization approach acting on the frequency spectra of generated images has been proposed, which supports the training process by penalizing whenever a generated image's spectrum deviates from the average spectrum of the real data. Our approach is related to (Durall, Keuper, and Keuper 2020). However, we argue that a pointwise regularization of all generated spectra w.r.t. the average spectrum of real images is suboptimal since it does not properly allow to learn the data distribution. Instead, we propose to use a discriminator on the power spectra in order to learn the generation of images according to both, the distribution of the real data in spatial as well as in frequency domain.

Contributions. We make the following contributions:

- We propose to learn to generate images with a higher fidelity to the real images' frequency distribution by employing a discriminator that acts on the frequency spectra.

- The proposed discriminator is efficient and modular and can be trained stably with different GAN losses.

- We propose a measure for the spectral distribution fidelity which allows to assess how well generated images can be distinguished from real ones by their frequency spectra.

- We show in various experiments that the proposed approach enables to generate images with highly realistic frequency spectra and therein outperforms the recent method from (Durall, Keuper, and Keuper 2020) without sacrificing image quality in terms of FID.
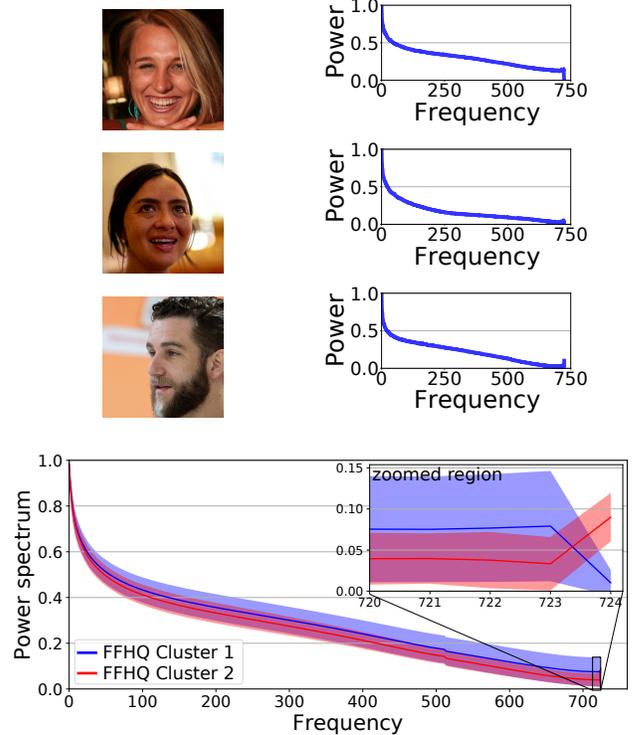

Figure 1: (top) The frequency profiles from the FFHQ images are diverse, suggesting that their distribution is not unimodal. (bottom) Average spectral profiles of the real data (FFHQ) after clustering with k-means (k=2). In the highest frequencies, the profiles can be well separated in two clusters.

Spectral Properties of Image Generation

Generative neural network architectures such as GANs generate high-dimensional outputs (i.e. high-resolution images) from low dimensional latent space samples. Therefore, they rely on stepwise up-scaling mechanisms which successively increase the output resolution, followed by convolutional layers. Such up-sampling can be done for example using "bed of nails", nearest neighbor or bilinear interpolation, all of which have different effects on the properties of the resulting up-sampled feature map or image. The spectral properties of an image I can be analyzed by its discrete Fourier transform

$$\hat{I}(k, \ell) = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} e^{-2\pi i \cdot \frac{m \cdot k}{M}} e^{-2\pi i \cdot \frac{n \cdot \ell}{N}} \cdot I(m, n), \quad (1)$$

$$\text{for} \quad k = 0, \ldots, M-1, \quad \ell = 0, \ldots, N-1,$$

which transforms a 2D image into a 2D array of its spatial frequencies. During up-sampling, an images frequency spectrum is altered depending on the up-sampling method. While bilinear interpolation results in smooth images with few high frequency components, the more commonly used "bed of nails" interpolation up-sampling, which fills the missing values with zeros, initializes the up-sampled image

(or feature map) with many high frequency components. In (Durall, Keuper, and Keuper 2020), a theoretic analysis of the effect of upsampling using bed of nails interpolation is provided, which we summarize below.

Spectral Effects of Up-Sampling

For simplicity, (Durall, Keuper, and Keuper 2020) consider in their analysis the case of one-dimensional signals, which generalizes to higher dimensions. For a signal a the discrete Fourier Transform $\hat{a}$ is given by

$$\hat{a}_k = \sum_{j=0}^{N-1} e^{-2\pi i \cdot \frac{jk}{N}} \cdot a_j, \quad \text{for} \quad k = 0, \dots, N-1. \quad (2)$$

Increasing a's spatial resolution by factor 2 results in

$$\hat{a}_{\bar{k}}^{up} = \sum_{j=0}^{2 \cdot N - 1} e^{-2\pi i \cdot \frac{j\bar{k}}{2 \cdot N}} \cdot a_j^{up}$$

$$= \sum_{j=0}^{N-1} e^{-2\pi i \cdot \frac{2 \cdot j\bar{k}}{2 \cdot N}} \cdot a_j + \sum_{j=0}^{N-1} e^{-2\pi i \cdot \frac{2 \cdot (j+1)\bar{k}}{2 \cdot N}} \cdot b_j, \quad (3)$$

for $\bar{k} = 0, \dots, 2N - 1$, where $b_j = 0$ for "bed of nails" interpolation (and $b_j = a_j$ for nearest neighbor interpolation). The first term in Eq. (3) is similar to the original Fourier Transform while the second term is zero for $b_j = 0$. It can be seen that if the spatial resolution is increased by a factor of 2 the frequency axes are scaled by a factor of $\frac{1}{2}$. From sampling theoretic considerations, it is (Durall, Keuper, and Keuper 2020)

$$(3) = \sum_{j=0}^{2 \cdot N - 1} e^{-2\pi i \cdot \frac{j\bar{k}}{2 \cdot N}} \cdot \sum_{t=-\infty}^{\infty} a_j^{up} \cdot \delta(j - 2t). \quad (4)$$

The point-wise multiplication with the Dirac impulse comb removes exactly the values for which $a^{up} = 0$. In order to apply the convolution theorem (Katznelson 2004), one has to assume a being a periodic signal. Then, it is

$$\hat{a}_{\bar{k}}^{up} = \frac{1}{2} \cdot \sum_{t=-\infty}^{\infty} \left(\sum_{j=-\infty}^{\infty} e^{-2\pi i \cdot \frac{j\bar{k}}{2 \cdot N}} a_j^{up} \right) \left(\bar{k} - \frac{t}{2} \right)$$

$$\overset{(3)}{=} \frac{1}{2} \cdot \sum_{t=-\infty}^{\infty} \left(\sum_{j=-\infty}^{\infty} e^{-2\pi i \cdot \frac{j\bar{k}}{N}} \cdot a_j \right) \left(\bar{k} - \frac{t}{2} \right). \quad (5)$$

Thus, the frequency spectrum $\hat{a}^{up}$ will contain replica of the frequency spectrum of a. More precisely, all frequencies beyond $\frac{N}{2}$ are up-sampling artifacts which can only be removed if the up-sampled signal is smoothed appropriately.

In addition to these theoretic considerations (Durall, Keuper, and Keuper 2020) also show practically that the correction of the resulting spectra is not possible with the commonly used 3×3 convolutional filters.

Analysis of Real Data Distribution

In order to analyze the distributions of the real as well as the generated images' frequency spectra, we consider an aggregate representation as done in (Durall, Keuper, and Keuper
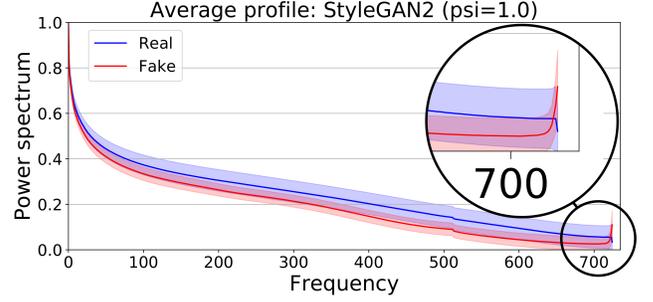

Figure 2: Average spectral profiles of real data (FFHQ) and data generated by StyleGAN2 (Karras et al. 2019). High frequency components in the generated images indicate grid artifacts or noise, which was not removed by the discriminator.

2020). We compute the magnitude of the 2D spectral frequencies and integrate over the resulting 2D array for every radius to get a one-dimensional profile of the power spectrum, called *azimuthal integral*

$$AI(\omega_k) = \int_0^{2\pi} \| \hat{I} \left(\omega_k \cdot \cos(\phi), \omega_k \cdot \sin(\phi) \right) \|^2 \mathrm{d}\phi$$

$$\text{for} \quad k = 0, \dots, M/2 - 1, \quad (6)$$

assuming square images I. As pointed out in (Durall, Keuper, and Keuper 2020) this notation is obviously abusive for $\hat{I}$ being discrete. Practically, the integral is implemented as a sum over interpolated values. See Fig. 1(top) for some examples of real images from the Flickr-Faces-HQ dataset (FFHQ) (Karras, Laine, and Aila 2019a) and their corresponding frequency profiles computed using Eq. (6).

From these examples, one can see that the frequency profiles of the images are diverse and vary significantly, especially in the high frequency regime. Fig. 1(bottom) shows the average frequency profiles after clustering the images of FFHQ by the magnitude of their highest frequency. The profiles can be well clustered into two groups just by looking at the highest frequency, which indicates that the frequency distribution of the real data can not be well approximated by a uni-modal Gaussian distribution as done in (Durall, Keuper, and Keuper 2020). Since the true distribution is unknown, we argue that a suitable way of generating images with a higher spectral fidelity is to use an adversarial approach. Our model therefore contains an additional discriminator, taking as input the frequency profile of real and generated images.

GAN Evaluation in the Frequency Domain

In general, the evaluation of the quality of generated images by GANs is highly subjective. Therefore, (Heusel et al. 2017) introduced the *Fréchet Inception Distance* (FID) to provide a method of comparing different image generating models. Since its proposal, this quality measure is widely adopted as one of the key indicators to proof the qualitative performance of GANs. At the time of writing, StyleGAN2 (Karras et al. 2019) is the best performing model according

to FID, and trained on the FFHQ dataset. Subjectively, generated images by StyleGAN2 are hard to distinguish from the real images. This is reflected by a low FID score of 2.84 ± 0.03. However, recent advances in DeepFake detection (Durall et al. 2019; Durall, Keuper, and Keuper 2020; Frank et al. 2020) showed that it is possible to recognize such images using frequency domain representations. Figure 2 shows the averaged profiles of a real data distribution, represented by FFHQ, and the corresponding images drawn from the learned distribution by StyleGAN2. It is visible that the power spectrum is misaligned throughout most frequencies. Images produced by StyleGAN2 contain high frequency components, which can be observed by a rapid increase in the power spectrum of the last frequencies. Such behavior indicates the presence of grid artifacts and high frequency noise. Thus, we argue that generated images should not only be assessed by their FID score but also by their spectral distribution fidelity and propose a method to evaluate the distribution alignment in the frequency domain, as a supplementary metric. Our method transforms images into spectral profile vectors by azimuthal integration (Eq. (6)). Then, we train a Logistic Regression on the spectral profiles given by all images taken from the real distribution, and a corresponding number of generated images given by a model. We then translate the training accuracy into a score according to

$$\text{Cloaking Score} = 1 - 2 \cdot |\text{Accuracy} - 0.5|. \quad (7)$$

The *Cloaking Score* (CS) ranges in $[0, 1]$, where a score of 1.0 indicates perfect spectral alignment, and a score of 0.0 indicates that generated images can be linearly separated from real images in the spectral domain. For the Style-GAN2 (Karras et al. 2019) example in Figure 2, the CS is 0.042 (acc. 0.979) after 1 000 epochs, 0.022 (acc. 0.989) after 10 000 epochs, and 0.018 (acc. 0.991) after 40 000 training epochs. Since the CS evaluation method gives rise to a trade-off between preciseness and runtime, we settled for scores after 1 000 epochs. Runtime for $140k$ 64^2-sized images is around 3min, when all images are read from disk.

Learning to Generate Spectral Distributions

Generative adversarial networks are trained in a minimax game between generator and discriminator networks, where the discriminator wants to recognize generated images and the generator wants the discriminator to perform poorly, i.e. to generate images which the discriminator can not tell apart from real ones, leading to objectives such as (Goodfellow et al. 2014)

$$\min_G \max_D V(D, G) = \quad \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})}[\log(D(\mathbf{x}))] +$$
$$\mathbb{E}_{\mathbf{z} \sim p_{\mathbf{z}}(\mathbf{z})}[\log(1 - D(G(\mathbf{z})))]. \quad (8)$$

Durall, Keuper, and Keuper (2020) propose to add a penalty term, a *Spectral Regularization*, to the generator to reduce discrepancies between the real and generated spectral distributions. This regularization is implemented as a cross entropy loss on the AI (6) profiles of the generator output, to minimize the difference of the generated images' spectra to

$\mathscr{L}_{\text{DCGAN}}$	$-\mathbb{E}_{\mathbf{x}}[\log(D(\mathbf{x}))] - \mathbb{E}_{\hat{\mathbf{x}}}[\log(1 - D(\hat{\mathbf{x}}))]$
$\mathscr{L}_{\text{LSGAN}}$	$-\mathbb{E}_{\mathbf{x}}[(D(\mathbf{x}) - 1)^2] + \mathbb{E}_{\hat{\mathbf{x}}}[D(\hat{\mathbf{x}})^2]$
$\mathscr{L}_{\text{WGAN}}$	$-\mathbb{E}_{\mathbf{x}}[D(\mathbf{x})] + \mathbb{E}_{\hat{\mathbf{x}}}[D(\hat{\mathbf{x}})]$
$\mathscr{L}_{\text{WGAN-GP}}$	$\mathscr{L}_{\text{WGAN}}+$
	$\lambda\mathbb{E}_{\hat{\mathbf{x}}}[(\|\nabla D(\alpha\mathbf{x} + (1 - \alpha\hat{\mathbf{x}}))\|_2 - 1)^2]$

Table 1: Investigated Discriminator Losses.

the mean spectrum of real images. Hence, the generator tries to minimize this penalty by forcing each image to reflect the same, average profile. In our experiments, this leads to an unstable training progress (see supplementary material). Out of 5 runs, only 3 models were able to reach 500 epochs without degenerating due to mode collapse. As we have shown above, the real spectral distribution is not a uni-modal Gaussian distribution around the average profile. Instead, there are certain characteristics that the generator needs to be able to learn. The regularizer by (Durall, Keuper, and Keuper 2020) seems suboptimal in this respect. Therefore, we argue that instead of learning one average profile, the generator should be taught to generate images according to the real data distribution in spatial as well as in spectral domain.

We propose to use a second discriminator network (D_F) for this purpose (see Figure 3). While one could obviously consider to use full 2D power spectra and a convolutional architecture as in the original discriminator, we argue that an additional discriminator should be as lightweight and modular as possible. The proposed frequency spectrum discriminator D_F takes as input real or generated images which then go through a spectral transformation layer ψ that computes the magnitude of their Fourier transform. Afterwards, the azimuthal integral (Eq. (6)) is computed, projecting the 2D spectrum onto a 1D vector. Then, we add a fully connected layer with Sigmoid activation function, which is trained to discriminate between real and generated images.

Such a simple discriminator can in principle be trained with any of the commonly used GAN loss functions. To investigate the dependency of the discriminator performance on the employed loss, we consider models with the commonly used loss functions in Tab. 1 and focus on simple architectures (i.e. DCGAN (Radford, Metz, and Chintala 2015), LSGAN (Mao et al. 2017), and Wasserstein GAN (WGAN) (Gulrajani et al. 2017a,b)).

We train both the spatial (D_S) and the frequency (D_F) discriminators by separate losses, applying the same loss function (e.g. BCE in case of DCGAN). For training the generator G, both resulting losses are combined by averaging (see Fig. 3). Additionally, we keep the proposed architectural suggestions by (Durall, Keuper, and Keuper 2020), namely increasing the last upconvolutional layer to 8×8, and adding an additional block of three 5×5 to the end of the generator network, which, as they argue empirically, is necessary to provide the network with the capacity to "repair" the spectral artifacts.

Implementation Details. The memory comsumption of the proposed discriminator depends on the image width
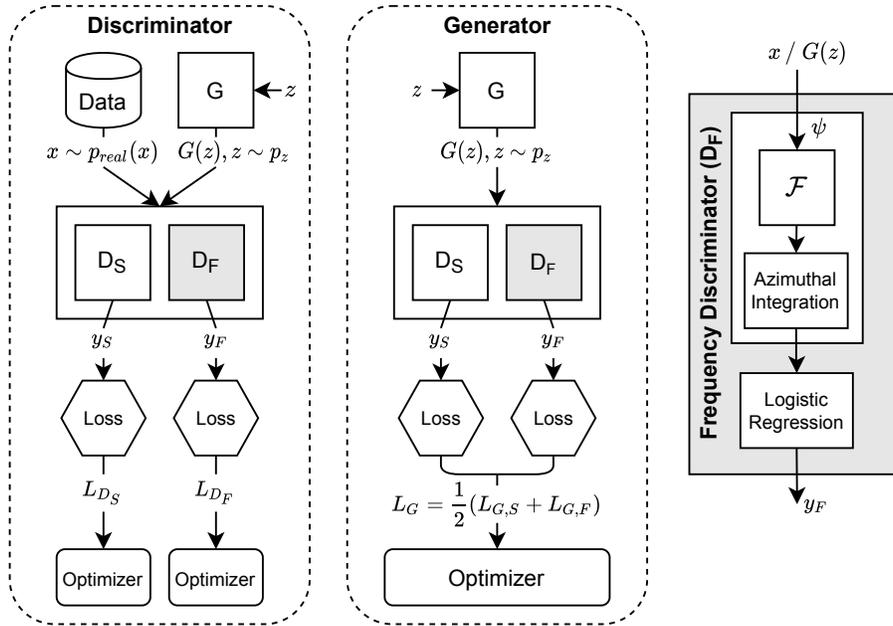
Figure 3: Training process of the proposed model. Losses are computed based on predictions representing the realness score of the given image in the spatial domain (y_S) as well as the frequency domain (y_F). Both spatial (D_S) and frequency (D_F) discriminators are trained separately by individual optimizer. When training generator G, both resulting losses are averaged.

Model	FID ↓	SD ↓	CS ↑
DCGAN$_{orig}$	17.749	1.510	**0.35**
DCGAN	**15.257**	**1.293**	0.16
LSGAN$_{orig}$	18.423	1.602	**0.28**
LSGAN	**15.518**	**0.468**	0.04

Table 2: Evaluation of the architectural change by Durall, Keuper, and Keuper (2020) on the image generation quality for DCGAN and LSGAN at 64^2 pixel resolution. In both cases, the changed up-sampling and additional convolutional layers lead to an improved FID and spectral difference, while the cloaking score is low.

and height n and adds only $\lceil n/\sqrt{2} \rceil + 1$ parameters. We evaluate it with different architectures. DCGAN, LSGAN and WGAN-GP were trained using Adam optimizer, for WGAN, we used RMSprop, with a learning rate of 0.0002 and a batch size of 128 for 500 epochs. In all cases, we apply the same loss function on both discriminators. Our implementation is publicly available[1].

Experiments

Experiments are conducted on the FFHQ dataset, which contains 70 000 high-quality face images at 1024×1024 resolution showing large variations in terms of age, ethnicity, and backgrounds. From this dataset, we downsample three versions in resolution 64×64, 128×128 and 256×256. We evaluate all experiments on 10k examples in terms of

---

[1]https://github.com/steffen-jung/SpectralGAN

FID and the proposed cloaking score (CS). If the distribution of the real data is matched in both spatial and frequency domain, the resulting FID should be low while the cloaking score should be high. Additionally, we report the sum of absolute differences between the average profile of 10k generated images with the average profile of all real images, denoted as *spectral difference* (SD). This measure is similar to the optimization target in (Durall, Keuper, and Keuper 2020).

First, we validate the impact of adding additional convolutional layers and increasing the filter size to 5×5 as suggested by Durall, Keuper, and Keuper (2020) on DCGAN and LSGAN (see Tab. 2). In both cases, the modification leads to an improved FID and a lower spectral difference. Yet, the cloaking score is even lower than the one of the baseline approach, indicating that the generated distributions are still linearly separable from their integrated frequency spectra (AI, Eq. (6)). This leads to two observations: (1) Increasing the amount of convolutions at the highest resolution can increase image quality of simple GAN methods, and (2) the conventional spatial discriminator does not provide the necessary signal to bring the spectral distributions closer together. From these observations, we continue with the modified networks to investigate whether the proposed spectral discriminator can provide the necessary training signal, and is able to use the network capacity, to assimilate the generated images' frequency spectra to the real distribution.

To evaluate the proposed spectral discriminator D_F in the context of diverse losses, we train variants of DCGAN (Radford, Metz, and Chintala 2015), LSGAN (Mao et al. 2017), WGAN (Arjovsky, Chintala, and Bottou 2017) and

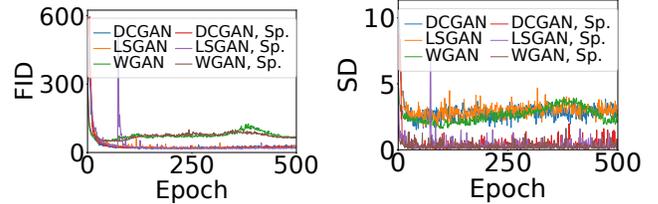| | | GAN Quality | | | Detection Accuracy | | | | |
| | | | | | Durall et al.* | | Durall et al. | | Wang et al. |
Size	Model	FID ↓	SD ↓	CS ↑	SVM ↓	LR ↓	SVM ↓	LR ↓	ACC ↓
64^2	DCGAN	**15.257**	1.293	0.16	0.89	0.89	0.89	0.89	0.81
	DCGAN (Durall et al.)	29.875	0.311	0.25	0.51	**0.49**	0.79	0.82	0.67
	DCGAN, Spectral	15.591	**0.042**	**0.84**	**0.50**	0.50	**0.59**	**0.58**	**0.66**
	LSGAN	15.518	0.468	0.04	0.94	0.94	0.94	0.94	0.67
	LSGAN, Spectral	**15.515**	**0.041**	**0.86**	**0.51**	0.50	**0.55**	**0.56**	**0.64**
	WGAN	**47.704**	1.291	0.01	0.99	0.99	0.99	0.99	0.79
	WGAN, Spectral	47.948	**0.029**	**0.85**	**0.50**	0.50	**0.62**	**0.65**	0.79
	WGAN-GP	**39.404**	0.575	0.18	0.92	0.94	0.92	0.94	0.76
	WGAN-GP, Spectral	39.441	**0.252**	**0.54**	**0.51**	0.51	**0.75**	**0.76**	**0.65**
128^2	DCGAN	19.908	1.579	0.00	0.99	0.99	0.99	0.99	0.82
	DCGAN (Durall et al.)	41.896	0.449	0.08	**0.51**	0.51	0.98	0.98	0.87
	DCGAN, Spectral	**19.898**	**0.087**	**0.72**	**0.51**	0.50	**0.66**	**0.69**	**0.81**
	LSGAN	20.850	3.786	0.01	0.99	0.99	0.99	0.99	0.81
	LSGAN, Spectral	**19.043**	**0.074**	**0.76**	**0.50**	0.50	**0.82**	**0.83**	**0.80**
	WGAN	**20.757**	1.996	0.01	0.99	0.99	0.99	0.99	0.83
	WGAN, Spectral	24.568	**0.136**	**0.66**	**0.52**	0.53	**0.61**	**0.60**	**0.81**
	WGAN-GP	47.630	2.349	0.02	0.99	0.99	0.99	0.99	0.96
	WGAN-GP, Spectral	**42.740**	**0.260**	**0.38**	**0.50**	0.50	**0.82**	**0.83**	**0.62**
256^2	DCGAN	**20.132**	9.826	0.00	1.00	1.00	1.00	1.00	0.96
	DCGAN (Durall et al.)	29.813	0.786	0.01	0.50	0.50	1.00	0.99	0.99
	DCGAN, Spectral	20.863	**0.229**	**0.50**	**0.50**	0.50	**0.89**	**0.93**	**0.85**
	LSGAN	19.893	5.955	0.00	1.00	1.00	1.00	1.00	0.85
	LSGAN, Spectral	**19.884**	**0.216**	**0.46**	**0.50**	0.50	**0.77**	**0.82**	**0.72**

Table 3: Evaluation of the proposed discriminator in terms of GAN quality measures and generated image detection scores.

| | | GAN Quality | | |
Size	Model	FID ↓	SD ↓	CS ↑
1024^2	StyleGAN2	**2.733**	32.250	0.09
	StyleGAN2, Spectral	3.326	**5.978**	**0.24**

Table 4: Evaluation of the finetuned StyleGAN2.

(a) The resulting FID scores are similar, and the training is stable. (b) With the proposed discriminator D_F, the spectral difference is considerably reduced.

Figure 4: FID and SD with our models trained on FFHQ64. *Sp.* indicates that D_F was used, and omitted otherwise.

WGAN GP (Gulrajani et al. 2017b) with and without D_F. While the modified DCGAN and LSGAN models can generate images at up to 256^2 pixel resolution, the training behavior of the Wasserstein GANs becomes easily unstable. Thus, we only consider them until 128^2 pixels resolution. Additionally, to evaluate scalability to high resolutions, we finetune a pretrained state-of-the-art model, StyleGAN2, after adding D_F to its training process.

The resulting FIDs over the training epochs as well as the spectral differences are displayed in Fig. 4a and Fig. 4b, respectively. It can be seen that the proposed discriminator has a relatively small impact on the FID during training while the spectral difference is significantly reduced by our method. The effect on the spectrum can be assessed in Tab. 3 (left block) and in Tab. 4. For all models and resolutions, the FID is on par with the plain version without additional discriminator while the spectral difference is considerably decreased and the cloaking score increased when D_F is added. The spectral regularization proposed by Durall, Keuper, and Keuper (2020) yields a significantly higher FID and lower cloaking score than the proposed approach.

In Tab. 3 (right block), we further evaluate the effect of the proposed discriminator on the detectability of the generated images using recent methods from Durall, Keuper, and Keuper (2020) and Wang et al. (2020). While Durall, Keuper, and Keuper (2020) leverage the 1D frequency spectra for the detection in a simple support vector machine (SVM) or logistic regression (LR) classifier and require retraining in every setting, Wang et al. (2020) train a CNN on the generated images using various data augmentations as a "universal" detector. We evaluate the method from Durall, Keuper, and Keuper (2020) in two scenarios: in the left column in Tab. 3 (right block) indicated with the asterix, we train on the data without spectral discriminator and evaluate in the transfer setting; in the right column, we train separately for
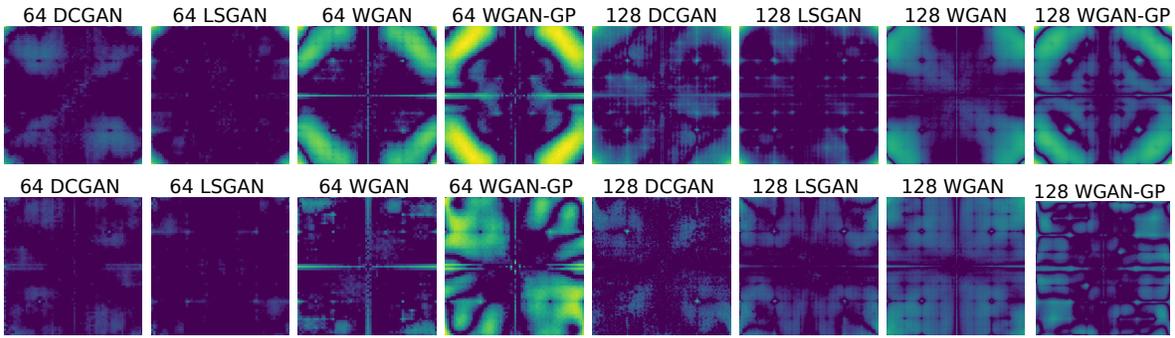
Figure 5: Mean absolute differences of the 2D power spectra of real and generated images. (Top) Differences for the original models. (Bottom) Differences for the models with spectral discriminator D_F. While the differences of the original models follow a characteristic pattern, our modified models show less specific differences and lower magnitudes.
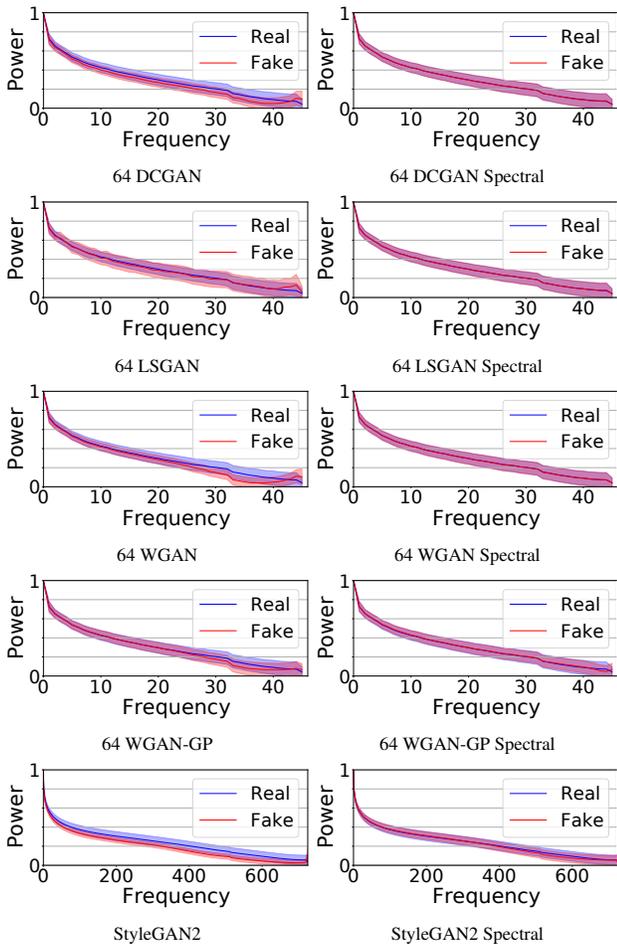

Figure 6: Resulting spectral profiles of experiments with our models trained on FFHQ64 and StyleGAN2. *Spectral* indicates that D_F was applied. With the spectral discriminator, the mean and standard deviation of the spectral profiles fit almost perfectly.

every setting. As can be seen, the detection is almost random in the transfer setting for the regularization approach by Durall, Keuper, and Keuper (2020) as well as for our approach in all tested GAN settings and resolutions while the generated images from (Durall, Keuper, and Keuper 2020) can still be detected with about 80% accuracy when a model is trained specifically on this data. With respect to the universal detector by Wang et al. (2020), both method decrease the detection accuracy significantly. This confirms that the removed frequency artifacts can in fact be perceived in the image domain and their removal is to be desired.

Figure 6 gives a visualization of the resulting average frequency profiles with our approach as an overlay on the average frequency of real images. With the proposed discriminator, the average frequencies fit almost perfectly. For the respective higher resolution experiments, as well as for example images generated by our approach see the supplementary material. Figure 5 shows the average differences of real and generated 2D power spectra in log-space (compare the visualization in (Wang et al. 2020)). It can be seen that all tested GANs show specific differences. When the proposed discriminator, acting on the 1D projection of the spectra, is employed, the differences become more diffuse and have lower magnitudes. This indicates that, on average, the generated images have less perceivable sampling artifacts.

Conclusion

In this paper, we proposed an adversarial image generation approach that enables to generate images with a significantly reduced amount of spectral artifacts. The proposed method employs a simple discriminator on the 1D projections of frequency spectra of real and generated images. Thus, the generator aims to match the real data distribution not only in the image domain but also in the frequency domain. Our approach is very lightweight and allows for a stable training, as we experimentally show for different loss functions and resolutions. It generates images that can not easily be distinguished from real ones by frequency artifacts and improves, in this respect, over the recent method by (Durall, Keuper, and Keuper 2020).

References

Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein gan. *arXiv preprint arXiv:1701.07875* .

Bai, Y.; Guo, Y.; Wei, J.; Lu, L.; Wang, R.; and Wang, Y. 2020. Fake Generated Painting Detection Via Frequency Analysis. In *2020 IEEE International Conference on Image Processing (ICIP)*, 1256–1260.

Brock, A.; Donahue, J.; and Simonyan, K. 2018a. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096* .

Brock, A.; Donahue, J.; and Simonyan, K. 2018b. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In *International Conference on Learning Representations (ICLR)*.

Choi, Y.; Choi, M.; Kim, M.; Ha, J.-W.; Kim, S.; and Choo, J. 2018. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 8789–8797.

Dai, B.; Fidler, S.; Urtasun, R.; and Lin, D. 2017. Towards diverse and natural image descriptions via a conditional gan. In *Proceedings of the IEEE International Conference on Computer Vision*, 2970–2979.

Donahue, J.; Krähenbühl, P.; and Darrell, T. 2016. Adversarial feature learning. *arXiv preprint arXiv:1605.09782* .

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 *CVPR*.

Durall, R.; Keuper, M.; Pfreundt, F.-J.; and Keuper, J. 2019. Unmasking DeepFakes with simple Features. *arXiv preprint arXiv:1911.00686* .

Frank, J.; Eisenhofer, T.; Schönherr, L.; Fischer, A.; Kolossa, D.; and Holz, T. 2020. Leveraging Frequency Analysis for Deep Fake Image Recognition. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, 3247–3258. PMLR.

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 (NeurIPS)*.

Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; and Courville, A. C. 2017a. Improved training of wasserstein gans. In *Advances in neural information processing systems (NeurIPS)*, 5767–5777.

Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; and Courville, A. C. 2017b. Improved training of wasserstein gans. In *Advances in neural information processing systems*, 5767–5777.

Gurumurthy, S.; Kiran Sarvadevabhatla, R.; and Venkatesh Babu, R. 2017. Deligan: Generative adversarial networks for diverse and limited data. In *Proceedings*

of the IEEE Conference on Computer Vision and Pattern Recognition*, 166–174.

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, X.; Liu, M.-Y.; Belongie, S.; and Kautz, J. 2018. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 172–189.

Iizuka, S.; Simo-Serra, E.; and Ishikawa, H. 2017. Globally and locally consistent image completion. *ACM Transactions on Graphics (ToG)* 36(4): 107.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1125–1134.

Karras, T.; Aila, T.; Laine, S.; and Lehtinen, J. 2017. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196* .

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

Karras, T.; Laine, S.; and Aila, T. 2019b. 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. 2019. Analyzing and improving the image quality of stylegan. *arXiv preprint arXiv:1912.04958* .

Katznelson, Y. 2004. *An introduction to harmonic analysis*. Cambridge University Press.

Kodali, N.; Abernethy, J.; Hays, J.; and Kira, Z. 2017. On convergence and stability of gans. *arXiv preprint arXiv:1705.07215* .

Mao, X.; Li, Q.; Xie, H.; Lau, R. Y.; Wang, Z.; and Paul Smolley, S. 2017. Least squares generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 2794–2802.

Mirza, M.; and Osindero, S. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784* .

Miyato, T.; Kataoka, T.; Koyama, M.; and Yoshida, Y. 2018. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957* .

Mo, S.; Cho, M.; and Shin, J. 2019. Instance-aware Image-to-Image Translation. In *International Conference on Learning Representations*. URL https://openreview.net/forum?id=ryxwJhC9YX.

Pathak, D.; Krahenbuhl, P.; Donahue, J.; Darrell, T.; and Efros, A. A. 2016. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2536–2544.

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

Reed, S.; Akata, Z.; Yan, X.; Logeswaran, L.; Schiele, B.; and Lee, H. 2016. Generative Adversarial Text to Image Synthesis. In *33rd International Conference on Machine Learning (ICML)*.

Wang, S.-Y.; Wang, O.; Zhang, R.; Owens, A.; and Efros, A. A. 2020. CNN-Generated Images Are Surprisingly Easy to Spot... for Now. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yu, N.; Davis, L.; and Fritz, M. 2019. Attributing Fake Images to GANs: Learning and Analyzing GAN Fingerprints. In *International Conference on Computer Vision (ICCV)*. URL https://publications.cispa.saarland/2965/.

Zhang, H.; Xu, T.; Li, H.; Zhang, S.; Wang, X.; Huang, X.; and Metaxas, D. N. 2017. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 5907–5915.

Zhang, H.; Xu, T.; Li, H.; Zhang, S.; Wang, X.; Huang, X.; and Metaxas, D. N. 2018. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE transactions on pattern analysis and machine intelligence* 41(8): 1947–1962.

Zhang, X.; Karaman, S.; and Chang, S.-F. 2019. Detecting and simulating artifacts in gan fake images. *arXiv preprint arXiv:1907.06515* .

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017a. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 2223–2232.

Zhu, J.-Y.; Zhang, R.; Pathak, D.; Darrell, T.; Efros, A. A.; Wang, O.; and Shechtman, E. 2017b. Toward multimodal image-to-image translation. In *Advances in Neural Information Processing Systems*, 465–476.