Focal Frequency Loss for Image Reconstruction and Synthesis

Liming Jiang[1] Bo Dai[1] Wayne Wu[2] Chen Change Loy[1]✉
[1]S-Lab, Nanyang Technological University [2]SenseTime Research
{liming002, bo.dai, ccloy}@ntu.edu.sg wuwenyan@sensetime.com

Abstract

Image reconstruction and synthesis have witnessed remarkable progress thanks to the development of generative models. Nonetheless, gaps could still exist between the real and generated images, especially in the frequency domain. In this study, we show that narrowing gaps in the frequency domain can ameliorate image reconstruction and synthesis quality further. We propose a novel focal frequency loss, which allows a model to adaptively focus on frequency components that are hard to synthesize by down-weighting the easy ones. This objective function is complementary to existing spatial losses, offering great impedance against the loss of important frequency information due to the inherent bias of neural networks. We demonstrate the versatility and effectiveness of focal frequency loss to improve popular models, such as VAE, pix2pix, and SPADE, in both perceptual quality and quantitative performance. We further show its potential on StyleGAN2.[1, 2]

1. Introduction

We have seen remarkable progress in image reconstruction and synthesis along with the development of generative models [19, 34, 15, 32, 56], and the progress continues with the emergence of various powerful deep learning-based approaches [31, 43, 44, 55]. Despite their immense success, one could still observe the gaps between the real and generated images in certain cases.

These gaps are sometimes manifested in the form of artifacts that are discernible. For instance, upsampling layers using transposed convolutions tend to produce checkerboard artifacts [42]. The gaps, in some other cases, may only be revealed through the frequency spectrum analysis. Recent studies [59, 72, 22] in media forensics have shown some notable periodic patterns in the frequency spectra of manipulated images, which may be consistent with artifacts in the spatial domain. In Figure 1, we show some paired examples of real images and the fake ones generated by typical generative models for image reconstruction and synthesis. It is observed that the frequency domain gap between the real and fake images may be a common issue for these methods, albeit in slightly different forms.

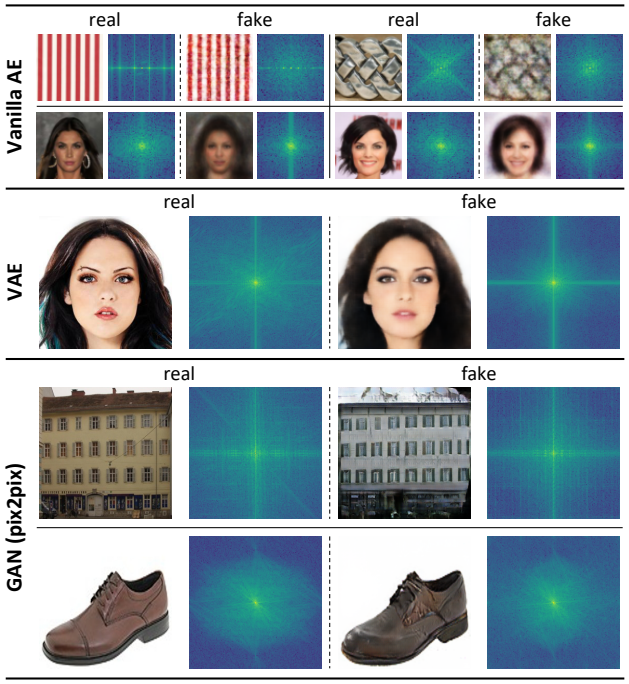

Figure 1. Frequency domain gaps between the real and the generated images by typical generative models in image reconstruction and synthesis. Vanilla AE [19] loses important frequencies, leading to blurry images (Row 1 and 2). VAE [34] biases to a limited spectrum region (Row 3), losing high-frequency information (outer regions and corners). Unnatural periodic patterns can be spotted on the spectra of images generated by GAN (pix2pix) [23] (Row 4), consistent with the observable checkerboard artifacts (zoom in for view). In some cases, a frequency spectrum region shift occurs to GAN-generated images (Row 5).

The observed gaps in the frequency domain may be imputed to some inherent bias of neural networks when applied to reconstruction and synthesis tasks. Fourier analysis highlights a phenomenon called *spectral bias* [48, 40, 54], a learning bias of neural networks towards low-frequency functions. Thus, generative models tend to eschew frequency components that are hard to synthesize, *i.e.*, hard frequencies, and converge to an inferior point. *F-Principle* [67] shows that the priority of fitting certain fre-

[1] GitHub: https://github.com/EndlessSora/focal-frequency-loss.
[2] Project page: https://www.mmlab-ntu.com/project/ffl/index.html.

quencies in a network is also different throughout the training, usually from low to high. Consequently, it is difficult for a model to maintain important frequency information as it tends to generate frequencies with a higher priority.

In this paper, we carefully study the frequency domain gap between real and fake images and explore ways to ameliorate reconstruction and synthesis quality by narrowing this gap. Existing methods [34, 23, 43] usually adopt pixel losses in the spatial domain, while spatial domain losses hardly help a network find hard frequencies and synthesize them, in that every pixel shares the same significance for a certain frequency. In comparison, we transform both the real and generated samples to their frequency representations using the standard discrete Fourier transform (DFT). The images are decomposed into sines and cosines, exhibiting periodic properties. Each coordinate value on the frequency spectrum depends on all the image pixels in the spatial domain, representing a specific spatial frequency. Explicitly minimizing the distance of coordinate values on the real and fake spectra can help networks easily locate difficult regions on the spectrum, *i.e*., hard frequencies.

To tackle these hard frequencies, inspired by hard example mining [11, 51] and focal loss [36], we propose a simple yet effective frequency-level objective function, named *focal frequency loss*. We map each spectrum coordinate value to a Euclidean vector in a two-dimensional space, with both the amplitude and phase information of the spatial frequency put under consideration. The proposed loss function is defined by the scaled Euclidean distance of these vectors by down-weighting easy frequencies using a dynamic spectrum weight matrix. Intuitively, the matrix is updated on the fly according to a non-uniform distribution on the current loss of each frequency during training. The model will then rapidly focus on hard frequencies and progressively refine the generated frequencies to improve image quality.

The main **contribution** of this work is a novel focal frequency loss that directly optimizes generative models in the frequency domain. We carefully motivate how a loss can be built on a space where frequencies of an image can be well represented and distinguished, facilitating optimization in the frequency dimension. We further explain the way that enables a model to focus on hard frequencies, which may be pivotal for quality improvement. Extensive experiments demonstrate the effectiveness of the proposed loss on representative baselines [19, 34, 23, 43], and the loss is complementary to existing spatial domain losses such as perceptual loss [27]. We further show the potential of focal frequency loss to improve state-of-the-art StyleGAN2 [31].

2. Related Work

Image reconstruction and synthesis. Autoencoders (AE) [19, 34] and generative adversarial networks (GAN) [15] are two popular models for image reconstruction and synthesis. The vanilla AE [19] aims at learning latent codes while reconstructing images. It is typically used for dimensionality reduction and feature learning. Autoencoders have been widely used to generate images since the development of variational autoencoders (VAE) [34, 33]. Their applications have been extended to various tasks, *e.g*., face manipulation [2, 1, 25, 24]. GAN [15, 41, 46], on the other hand, is extensively applied in face generation [29, 30, 31], image-to-image translation [23, 74, 7, 26], style transfer [37, 21], and semantic image synthesis [60, 43, 38]. Existing approaches usually apply spatial domain loss functions, *e.g*., perceptual loss [27], to improve quality while seldom consider optimization in the frequency domain. Spectral regularization [10] presents a preliminary attempt. Different from [27, 10], the proposed focal frequency loss dynamically focuses the model on hard frequencies by down-weighting the easy ones and ameliorates image quality through the frequency domain directly. Some concurrent works on image reconstruction and synthesis via the frequency domain include [4, 14, 28].

Frequency domain analysis of neural networks. In addition to the studies [48, 40, 54, 67] we discussed in the introduction, we highlight some recent works that analyze neural networks through the frequency domain. Using coordinate-based MLPs, Fourier features [54, 49] and positional encoding [40, 57] are adopted to recover missing high frequencies in single image regression problems. Besides, several studies have incorporated frequency analysis with network compression [16, 66, 68, 6, 17] and feature reduction [35, 61] to accelerate the training and inference of networks. The application areas of the frequency domain analysis have been further extended, including media forensics [59, 72, 22, 12], super-resolution [13, 63], generalization analysis [58, 20], magnetic resonance imaging [53], image rescaling [64], *etc*. Despite the wide exploration of various problems, improving reconstruction and synthesis quality via the frequency domain remains much less explored.

Hard example processing. Hard example processing is widely explored in object detection and image classification to address the class imbalance problem. A common solution is to use a bootstrapping technique called hard example mining [51, 11], where a representative method is online hard example mining (OHEM) [51]. The training examples are sampled following the current loss of each example to modify the stochastic gradient descent. The model is encouraged to learn hard examples more to boost performance. An alternative solution is focal loss [36], which is a scaled cross-entropy loss. The scaling factor down-weights the contribution of easy examples during training so that a model can focus on learning hard examples. The proposed focal frequency loss is inspired by these techniques.

3. Focal Frequency Loss

To formulate our method, we explicitly exploit the frequency representation of images (Section 3.1), facilitating

the network to locate the hard frequencies. We then define a frequency distance (Section 3.2) to quantify the differences between images in the frequency domain. Finally, we adopt a dynamic spectrum weighting scheme (Section 3.3) that allows the model to focus on the on-the-fly hard frequencies.

3.1. Frequency Representation of Images

In this section, we revisit and highlight several key concepts of the discrete Fourier transform. We demonstrate the effect of missing frequencies in the image and the advantage of frequency representation for locating hard frequencies.

Discrete Fourier transform (DFT) is a complex-valued function that converts a discrete finite signal into its constituent frequencies, *i.e.*, complex exponential waves. An image[3] can be treated as a two-dimensional discrete finite signal with only real numbers. Thus, to convert an image into its frequency representation, we perform the 2D discrete Fourier transform:

$$F(u,v) = \sum_{x=0}^{M-1}\sum_{y=0}^{N-1} f(x,y)\cdot e^{-i2\pi\left(\frac{ux}{M}+\frac{vy}{N}\right)}, \quad (1)$$

where the image size is $M \times N$; (x, y) denotes the coordinate of an image pixel in the spatial domain; $f(x, y)$ is the pixel value; (u, v) represents the coordinate of a spatial frequency on the frequency spectrum; $F(u, v)$ is the complex frequency value; e and i are Euler's number and the imaginary unit, respectively. Following Euler's formula:

$$e^{i\theta} = \cos\theta + i\sin\theta, \quad (2)$$

the natural exponential function in Eq. (1) can be written as:

$$e^{-i2\pi\left(\frac{ux}{M}+\frac{vy}{N}\right)} = \cos 2\pi\left(\frac{ux}{M}+\frac{vy}{N}\right) - i\sin 2\pi\left(\frac{ux}{M}+\frac{vy}{N}\right). \quad (3)$$

According to Eq. (1) and Eq. (3), the image is decomposed into orthogonal sine and cosine functions, constituting the imaginary and the real part of the frequency value, respectively, after applied 2D DFT. Each sine or cosine can be regarded as a binary function of (x, y), where its angular frequency is decided by the spectrum position (u, v). The mixture of these sines and cosines provides both the horizontal and vertical frequencies of an image. Therefore, spatial frequency manifests as the 2D sinusoidal components in the image. The spectrum coordinate (u, v) also represents the angled direction of a spatial frequency (visualizations can be found in the *supplementary material*), and $F(u, v)$ shows the "response" of the image to this frequency. Due to the periodicity of trigonometric functions, the frequency representation of an image also acquires periodic properties.

Note that in Eq. (1), $F(u, v)$ is the sum of a function that traverses every image pixel in the spatial domain, hence a specific spatial frequency on the spectrum depends on all the image pixels. For an intuitive visualization, we suppress the *single* center point (the lowest frequency) of the spectrum (Column 2 of Figure 2), leading to *all* the image pixels being affected. To further ascertain the spatial frequency at the different regions on the spectrum, we perform some other standard bandlimiting operations and visualize their physical meanings in the spatial domain (Figure 2). A low-pass filter (Column 3), *i.e.*, missing high frequencies, causes blur and typical ringing artifacts. A high-pass filter (Column 4), *i.e.*, missing low frequencies, tends to retain edges and boundaries. Interestingly, a simple band-stop filter (Column 5), *i.e.*, missing certain frequencies, produces visible common checkerboard artifacts (zoom in for view).

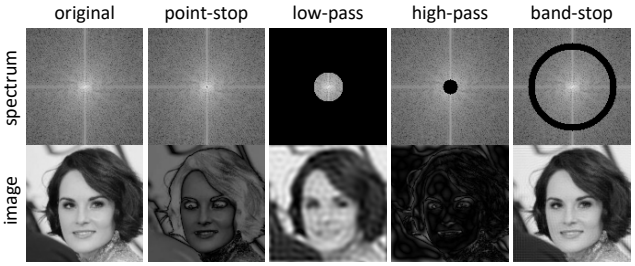

Figure 2. Standard bandlimiting operations on the frequency spectrum with the origin (low frequencies) center shifted and respective images in the spatial domain. These manual operations can be regarded as a simulation to show the effect of missing frequencies.

Observably, the losses of different regions on the frequency spectrum correspond to different artifacts on the image. One may deduce that compensating for these losses may reduce artifacts and improve image reconstruction and synthesis quality. The analysis here shows the value of using the frequency representation of images for profiling and locating different frequencies, especially the hard ones.

3.2. Frequency Distance

To devise a loss function for the missing frequencies, we need a distance metric that quantifies the differences between real and fake images in the frequency domain. The distance has to be differentiable to support stochastic gradient descent. In the frequency domain, the data objects are different spatial frequencies on the frequency spectrum, appearing as different 2D sinusoidal components in an image. To design our frequency distance, we further study the real and imaginary part of the complex value $F(u, v)$ in Eq. (1).

Let $R(u, v) = a$ and $I(u, v) = b$ be the real and the imaginary part of $F(u, v)$, respectively. $F(u, v)$ can be rewritten as:

$$F(u,v) = R(u,v) + I(u,v)\,i = a + bi. \quad (4)$$

According to the definition of 2D discrete Fourier transform, there are two key elements in $F(u, v)$.The first el-

[3] For simplicity, the formulas in this section are applied to gray-scale images, while the extension to color images is straightforward by processing each channel separately in the same way.

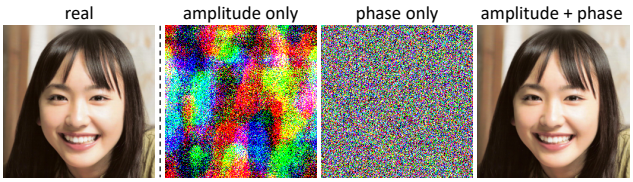

Figure 3. The necessity of both amplitude and phase information for a frequency distance verified by *single-image reconstruction*. "Amplitude/phase only" means solely applying Eq. (5)/(6) to calculate the distance between the real and reconstructed images.

ement is *amplitude*, which is defined as:

$$|F(u,v)| = \sqrt{R(u,v)^2 + I(u,v)^2} = \sqrt{a^2+b^2}. \quad (5)$$

Amplitude manifests the energy, *i.e.*, how strongly an image responds to the 2D sinusoidal wave with a specific frequency. We typically show the amplitude as an informative visualization of the frequency spectrum (*e.g.*, Figure 1 and 2). The second element is *phase*, which is written as:

$$\angle F(u,v) = \arctan\left(\frac{I(u,v)}{R(u,v)}\right) = \arctan\frac{b}{a}. \quad (6)$$

Phase represents the shift of a 2D sinusoidal wave from the wave with the origin value (the beginning of a cycle).

A frequency distance should consider both the amplitude and the phase as they capture different information of an image. We show a single-image reconstruction experiment in Figure 3. Merely minimizing the amplitude difference returns a reconstructed image with irregular color patterns. Conversely, using only the phase information, the synthesized image resembles a noise. A faithful reconstruction can only be achieved by considering both amplitude and phase.

Our solution is to map each frequency value to a Euclidean vector in a two-dimensional space (*i.e.*, a plane). Following the standard definition of a complex number, the real and imaginary parts correspond to the x-axis and y-axis, respectively. Let $F_r(u,v) = a_r + b_r i$ be the spatial frequency value at the spectrum coordinate (u,v) of the real image, and the corresponding $F_f(u,v) = a_f + b_f i$ with the similar meaning w.r.t. the fake image. We denote $\vec{r_r}$ and $\vec{r_f}$ as two respective vectors mapped from $F_r(u,v)$ and $F_f(u,v)$ (see Figure 4). Based on the definition of amplitude and phase, we note that the vector magnitude $|\vec{r_r}|$ and $|\vec{r_f}|$ correspond to the amplitude, and the angle θ_r and θ_f correspond to the phase. Thus, the frequency distance corresponds to the distance between $\vec{r_r}$ and $\vec{r_f}$, which considers both the vector magnitude and angle. We use the (squared) Euclidean distance for a single frequency:

$$d(\vec{r_r}, \vec{r_f}) = \|\vec{r_r} - \vec{r_f}\|_2^2 = |F_r(u,v) - F_f(u,v)|^2. \quad (7)$$

The frequency distance between the real and fake images can be written as the average value:

$$d(F_r, F_f) = \frac{1}{MN}\sum_{u=0}^{M-1}\sum_{v=0}^{N-1}|F_r(u,v) - F_f(u,v)|^2. \quad (8)$$

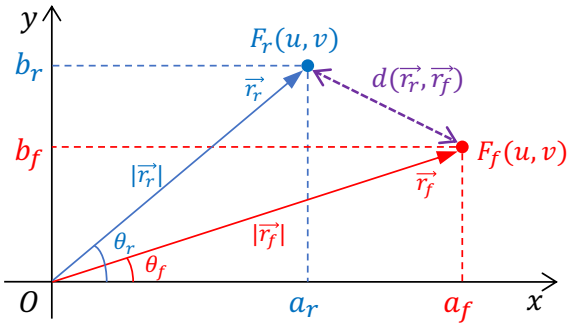

Figure 4. Frequency distance between $\vec{r_r}$ and $\vec{r_f}$ mapped from two corresponding real and fake frequency values $F_r(u,v)$ and $F_f(u,v)$ at the spectrum position (u,v). The Euclidean distance (purple line) is used, considering both the amplitude (magnitude $|\vec{r_r}|$ and $|\vec{r_f}|$) and phase (angle θ_r and θ_f) information.

3.3. Dynamic Spectrum Weighting

The frequency distance we defined in Eq. (8) quantitatively compares the real and fake images in the frequency domain. However, a direct use of Eq. (8) as a loss function is not helpful in coping with hard frequencies since the weight of each frequency is identical. A model would still bias to easy frequencies due to the inherent bias.

Inspired by hard example mining [11, 51] and focal loss [36], we formulate our method to focus the training on the hard frequencies. To implement this, we introduce a spectrum weight matrix to down-weight the easy frequencies. The spectrum weight matrix is dynamically determined by a non-uniform distribution on the current loss of each frequency during training. Each image has its own spectrum weight matrix. The shape of the matrix is the same as that of the spectrum. The matrix element $w(u,v)$, *i.e.*, the weight for the spatial frequency at (u,v), is defined as:

$$w(u,v) = |F_r(u,v) - F_f(u,v)|^\alpha, \quad (9)$$

where α is the scaling factor for flexibility ($\alpha = 1$ in our experiments). We further normalize the matrix values into the range $[0,1]$, where the weight 1 corresponds to the currently most lost frequency, and the easy frequencies are down-weighted. The gradient through the spectrum weight matrix is locked, so it only serves as the weight for each frequency.

By performing the Hadamard product for the spectrum weight matrix and the frequency distance matrix, we have the *full form* of the focal frequency loss (FFL):

$$\text{FFL} = \frac{1}{MN}\sum_{u=0}^{M-1}\sum_{v=0}^{N-1} w(u,v)\,|F_r(u,v) - F_f(u,v)|^2. \quad (10)$$

The focal frequency loss can be seen as a weighted average of the frequency distance between the real and fake images. It focuses the model on synthesizing hard frequencies by down-weighting easy frequencies. Besides, the focused region is updated on the fly to complement the immediate

hard frequencies, thus progressively refining the generated images and being adaptable to different methods.

In practice, to apply the proposed focal frequency loss to a model, we first transform both the real and fake images into their frequency presentations using the 2D DFT. We then perform the orthonormalization for each frequency value $F(u, v)$, *i.e.*, dividing it by $\sqrt{MN}$, so that the 2D DFT is unitary to ensure a smooth gradient. Finally, we employ Eq. (10) to calculate the focal frequency loss. We note that the exact form of focal frequency loss is not crucial. Some studies on the loss variants are provided in the *supplementary material*.

4. Experiments

4.1. Settings

Baselines. We start from image reconstruction by vanilla AE [19] (*i.e.*, a simple 2-layer MLP) and VAE [34] (*i.e.*, CNN-based). We then study unconditional image synthesis using VAE, *i.e.*, generating images from the Gaussian noise. Besides, we also investigate conditional image synthesis using GAN-based methods. Specifically, we select two typical image-to-image translation approaches, *i.e.*, pix2pix [23] and SPADE [43]. We further explore the potential of focal frequency loss (FFL) on state-of-the-art StyleGAN2 [31]. In addition, we compare FFL with relevant losses [27, 10]. The implementation details are provided in the *supplementary material*.

Datasets. We use a total of seven datasets. The datasets vary in types, sizes, and resolutions. For vanilla AE, we exploit the Describable Textures Dataset (DTD) [8] and CelebA [39]. For VAE, we use CelebA and CelebA-HQ [29] with different resolutions. For pix2pix, we utilize the officially prepared CMP Facades [47] and edges $\rightarrow$ shoes [69] datasets. For SPADE, we select two challenging datasets, *i.e.*, Cityscapes [9] and ADE20K [73]. For StyleGAN2, we reuse CelebA-HQ. Please refer to the *supplementary material* for the dataset details.

Evaluation metrics. To evaluate frequency domain difference, we introduce a frequency-level metric, named Log Frequency Distance (LFD), which is defined by a modified version of Eq. (8):

$$\text{LFD} = \log\left[\frac{1}{MN}\left(\sum_{u=0}^{M-1}\sum_{v=0}^{N-1}\left|F_r(u,v) - F_f(u,v)\right|^2\right) + 1\right], \tag{11}$$

where the logarithm is only used to scale the value into a reasonable range. A lower LFD is better. Note that LFD is a full reference metric (*i.e.*, requiring the ground truth image), so we use it in the reconstruction tasks.

Besides, we integrate the evaluation protocols from prior works [40, 3, 43, 26]. Specifically, we employ FID (lower is better) [18] for all tasks. For the reconstruction tasks of vanilla AE and VAE, we use PSNR (higher is better), SSIM (higher is better) [62], and LPIPS (lower is better) [71] in addition to LFD and FID. For the synthesis tasks of VAE, pix2pix, and StyleGAN2, we apply IS (higher is better) [50] in addition to FID. For SPADE (task-specific method for semantic image synthesis), besides FID, we follow their paper [43] to use mIoU (higher is better) and pixel accuracy (accu, higher is better) for the segmentation performance of synthesized images. We use DRN-D-105 [70] for Cityscapes and UperNet101 [65] for ADE20K.

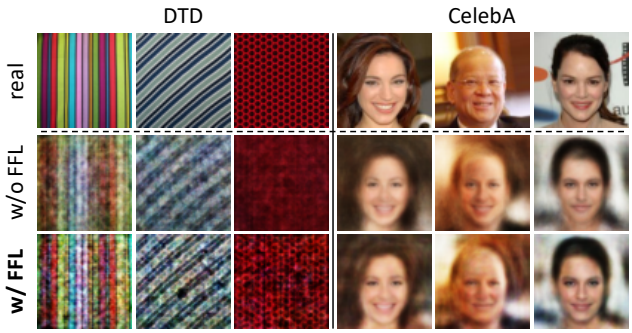

Figure 5. **Vanilla AE image reconstruction** results on the **DTD** (64×64) and **CelebA** (64×64) datasets.

Table 1. The PSNR (higher is better), SSIM (higher is better), LPIPS (lower is better), FID (lower is better) and LFD (lower is better) scores for the **vanilla AE image reconstruction** trained with/without the focal frequency loss (FFL).

Dataset	FFL	PSNR↑	SSIM↑	LPIPS↓	FID↓	LFD↓
DTD	w/o	20.133	**0.407**	0.414	246.870	14.764
	w/	**20.151**	0.400	**0.404**	**240.373**	**14.760**
CelebA	w/o	20.044	0.568	0.237	97.035	14.785
	w/	**21.703**	**0.642**	**0.199**	**83.801**	**14.403**

4.2. Results and Analysis

Vanilla AE. The results of vanilla AE [19] image reconstruction are shown in Figure 5. On DTD, without the focal frequency loss (FFL), the vanilla AE baseline synthesizes blurry images, which lack sufficient texture details and only contain some low-frequency information. With FFL, the reconstructed images become clearer and show more texture details. The results on CelebA show that FFL improves a series of quality problems, *e.g.*, face blur (Column 4), identity shift (Column 5), and expression loss (Column 6).

The quantitative evaluation results are presented in Table 1. Adding the proposed FFL to the vanilla AE baseline leads to a performance boost in most cases on the DTD and CelebA datasets w.r.t. five evaluation metrics. We note that the performance boost on CelebA is larger, indicating the effectiveness of FFL for the natural images.

VAE. The results of VAE [34] image reconstruction and unconditional image synthesis on CelebA are shown in Figure 6. For reconstruction, FFL helps the VAE model better retain the image clarity (Column 1), expression (Column 2), and skin color (Column 3). The unconditional synthesis results (Column 4, 5, 6) show that the quality of generated images is improved after applying FFL. The generated faces become clearer and gain more texture details. For a higher

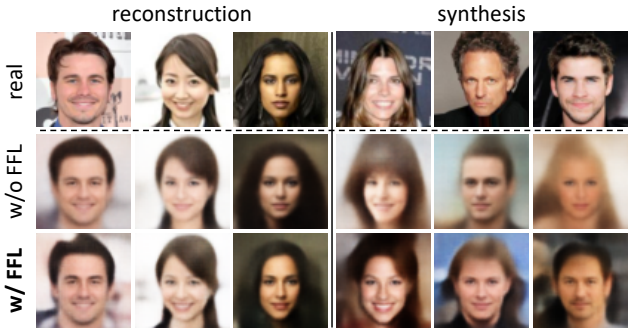

Figure 6. **VAE image reconstruction** and **unconditional image synthesis** results on the **CelebA** (64×64) dataset.

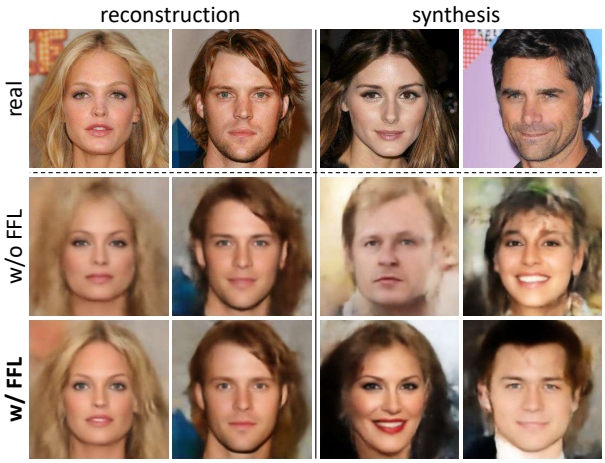

Figure 7. **VAE image reconstruction** and **unconditional image synthesis** results on the **CelebA-HQ** (256×256) dataset.

Table 2. The PSNR (higher is better), SSIM (higher is better), LPIPS (lower is better), FID (lower is better) and LFD (lower is better) scores for the **VAE image reconstruction** trained with/without the focal frequency loss (FFL).

Dataset	FFL	PSNR↑	SSIM↑	LPIPS↓	FID↓	LFD↓
CelebA	w/o	19.961	0.606	0.217	69.900	14.804
	w/	**22.954**	**0.723**	**0.143**	**49.689**	**14.115**
CelebA-HQ	w/o	21.310	0.616	0.367	71.081	17.266
	w/	**22.253**	**0.637**	**0.344**	**59.470**	**17.049**

Table 3. The FID (lower is better) and IS (higher is better) scores for the **VAE unconditional image synthesis** trained with/without the focal frequency loss (FFL).

Dataset	FFL	FID↓	IS↑
CelebA	w/o	80.116	1.873
	w/	**71.050**	**2.010**
CelebA-HQ	w/o	93.778	2.057
	w/	**84.472**	**2.060**

resolution, we present the VAE reconstruction and synthesis results on CelebA-HQ in Figure 7. By adding FFL to the VAE baseline, the reconstructed images keep more original image information, *e.g.*, mouth color (Column 2) and opening angle (Column 1). Besides, high-frequency details on the hair are clearly enhanced (Column 1). For unconditional image synthesis, FFL helps reduce artifacts and ameliorates the perceptual quality of synthesized images.

The quantitative test results of VAE image reconstruction are shown in Table 2. Adding FFL to the VAE baseline achieves better performance w.r.t. all the metrics. Besides, both FID and IS are better in the unconditional image synthesis task (Table 3), indicating that the generated images are clearer and more photorealistic. The results suggests the effectiveness of the focal frequency loss in helping VAE to improve image reconstruction and synthesis quality.

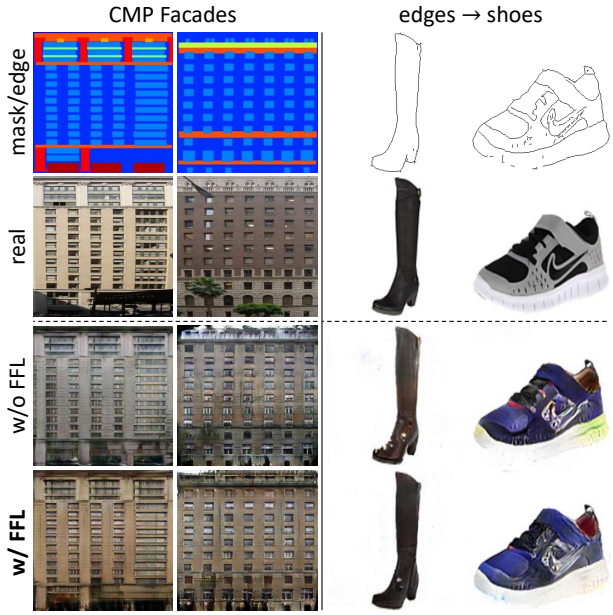

Figure 8. **pix2pix image-to-image translation** results on **CMP Facades** (256×256) and **edges → shoes** (256×256) datasets.

Table 4. The FID (lower is better) and IS (higher is better) scores for the **pix2pix image-to-image translation** trained with/without the focal frequency loss (FFL).

Dataset	FFL	FID↓	IS↑
CMP Facades	w/o	128.492	1.571
	w/	**123.773**	**1.738**
edges → shoes	w/o	80.279	2.674
	w/	**74.359**	**2.804**

pix2pix. For conditional image synthesis, the results of pix2pix [23] image-to-image translation (I2I) are shown in Figure 8. On CMP Facades, FFL improves the image synthesis quality of pix2pix by reducing unnatural colors (Column 1) or the black artifacts on the building (Column 2). Meanwhile, the semantic information alignment with the mask becomes better after applying FFL. For the edges → shoes translation, pix2pix baseline sometimes introduces colored checkerboard artifacts to the white background (Column 3, zoom in for view). Besides, atypical colors appear in certain cases (Column 4). In comparison, the model trained with FFL yields fewer artifacts.

The quantitative evaluation results of pix2pix image-to-image translation are shown in Table 4. FFL contributes to a performance boost on both of the two datasets. The results of the pix2pix baseline show the adaptability of the focal frequency loss for the image-to-image translation problem.

SPADE. We further explore semantic image synthesis (*i.e.*,

Figure 9. **StyleGAN2 unconditional image synthesis** results (without truncation) and the mini-batch average spectra (adjusted to better contrast) on the **CelebA-HQ** (256×256) dataset.

Figure 10. **SPADE semantic image synthesis** results on the **Cityscapes** (512×256) and **ADE20K** (256×256) datasets.

synthesizing a photorealistic image from a semantic segmentation mask) on more challenging datasets. The results of SPADE [43] are shown in Figure 10. In the street scene of Cityscapes (Column 1), SPADE baseline distorts the car and road, missing some important details (*e.g.*, road line). The model trained with FFL demonstrates better perceptual quality for these details. In the outdoor scene of ADE20K (Column 2), applying FFL to SPADE boosts its ability to generate details on the buildings. Besides, for the ADE20K indoor images (Column 3), SPADE baseline produces some abnormal artifacts in certain cases. The model trained with the proposed FFL synthesizes more photorealistic images.

The quantitative test results are presented in Table 5 (the values used for comparison are taken from [43]). We compare SPADE trained with/without FFL against a series of

Table 5. The mIoU (higher is better), pixel accuracy (accu, higher is better) and FID (lower is better) scores for the **SPADE semantic image synthesis** trained with/without the focal frequency loss (FFL) compared to a series of task-specific methods.

Method	Cityscapes			ADE20K		
	mIoU↑	accu↑	FID↓	mIoU↑	accu↑	FID↓
CRN [5]	52.4	77.1	104.7	22.4	68.8	73.3
SIMS [45]	47.2	75.5	**49.7**	N/A	N/A	N/A
pix2pixHD [60]	58.3	81.4	95.0	20.3	69.2	81.8
SPADE [43]	62.3	81.9	71.8	38.5	79.9	33.9
SPADE + FFL	**64.2**	**82.5**	59.5	**42.9**	**82.4**	**33.7**

Table 6. The FID (lower is better) and IS (higher is better) scores for the **StyleGAN2 unconditional image synthesis** trained with/without the focal frequency loss (FFL).

Dataset	FFL	FID↓	IS↑
CelebA-HQ (256×256)	w/o	5.696	3.383
	w/	**4.972**	**3.432**

open-source task-specific semantic image synthesis methods [5, 45, 60]. SIMS [45] obtains the best FID but poor segmentation scores on Cityscapes in that it directly stitches the training image patches from a memory bank while not keeping the exactly consistent positions. Without modifying the SPADE network structure, training with FFL contributes a further performance boost, greatly outperforming the benchmark methods, which suggests the effectiveness of FFL for semantic image synthesis.

StyleGAN2. We apply FFL to the mini-batch average spectra of the real images and the generated images by the state-of-the-art unconditional image synthesis method, *i.e.*, StyleGAN2 [31], intending to narrow the frequency distribution gap and improve quality further. The results on CelebA-HQ (256×256) without truncation [30, 31] are shown in Figure 9. Although StyleGAN2 (w/o FFL) generates photorealistic images in most cases, some tiny artifacts can still be spotted on the background (Column 2 and 4) and face (Column 5). Applying FFL, such artifacts are reduced, ameliorating synthesis quality further. Observably, the frequency domain gaps between mini-batch average spectra are clearly mitigated by FFL (Column 8). Some higher-resolution re-

Table 7. **Comparison** of our focal frequency loss (FFL) **with relevant losses**, *i.e.*, perceptual loss (PL), spectral regularization (SpReg), and another transformation form for FFL, *i.e.*, discrete cosine transform (DCT), in different image reconstruction and synthesis tasks.

Method	VAE reconstruction (CelebA)					VAE synthesis (CelebA)		pix2pix I2I (edges → shoes)	
	PSNR↑	SSIM↑	LPIPS↓	FID↓	LFD↓	FID↓	IS↑	FID↓	IS↑
baseline	19.961	0.606	0.217	69.900	14.804	80.116	1.873	80.279	2.674
+ PL [27]	20.964	0.658	**0.143**	62.795	14.573	78.825	1.788	78.916	2.722
+ SpReg [10]	19.974	0.607	0.218	69.118	14.796	78.079	1.898	79.300	2.700
+ FFL (DCT)	22.677	0.711	0.150	51.536	14.179	71.827	1.932	79.045	2.754
+ FFL (Ours)	**22.954**	**0.723**	**0.143**	**49.689**	**14.115**	**71.050**	**2.010**	**74.359**	**2.804**

sults can be found in the *supplementary material*.

The quantitative results are reported in Table 6. FFL improves both FID and IS, in line with the visual quality enhancement. The results on StyleGAN2 show the potential of FFL to boost state-of-the-art baseline performance.

Comparison with relevant losses. For completeness and fairness, we compare the proposed focal frequency loss (FFL) with relevant loss functions that aim at improving image reconstruction and synthesis quality. Specifically, we select the widely used spatial-based method, *i.e.*, perceptual loss (PL) [27], which depends on high-level features from a pre-trained VGG [52] network. We also study the frequency-based approach, *i.e.*, spectral regularization (SpReg) [10], which is derived based on the azimuthal integration of the Fourier power spectrum. Besides, we further compare with another transformation form for FFL, *i.e.*, discrete cosine transform (DCT).

The comparison results are reported in Table 7. FFL outperforms the relevant approaches (*i.e.*, PL and SpReg) when applied to our baselines in different image reconstruction and synthesis tasks. It is noteworthy that FFL and PL are complementary, as shown by our previous experiments on SPADE, which also uses PL. Even if we replace DFT with DCT as the transformation form of FFL, the results are still better than previous methods. The performance is only slightly inferior to that obtained by FFL with DFT (*i.e.*, Eq. (10)). We deduce that the transformation form for FFL may be flexible. At this stage, DFT may be a better choice.

Ablation studies. We present ablation studies of each key component for FFL in Figure 11 and corresponding quantitative results in Table 8. For intuitiveness and simplicity, we use vanilla AE image reconstruction on CelebA for the evaluation.

The full FFL achieves the best performance. If we do not use the frequency representation of images (Section 3.1) and focus the model on hard pixels in the spatial domain, the synthesized images become more blurry. The quantitative results degrade. Discarding either the phase or amplitude information (Section 3.2) harms the metric performance vastly. Visually, using no phase information (amplitude only), the contour of reconstructed faces is retained, but the color is completely shifted. Without amplitude (phase only), the model cannot reconstruct the faces at all, and the full identity information is lost. This further verifies the necessity of considering both amplitude and phase information. Without focusing the model on the hard frequencies by the dynamic spectrum weighting (*i.e.*, directly using Eq. (8)), the results are visually similar to baseline, in line with our discussion in Section 3.3. The metrics decrease, being close to but slightly better than baseline, which may benefit from the frequency representation.

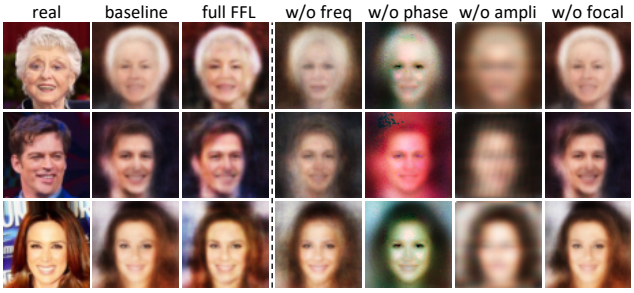

Figure 11. **Ablation studies** of each key component for the focal frequency loss (FFL), *i.e.*, frequency representation (freq), phase and amplitude (ampli) information, and dynamic spectrum weighting (focal) in the vanilla AE image reconstruction task on CelebA.

Table 8. The PSNR (higher is better), SSIM (higher is better), LPIPS (lower is better), FID (lower is better) and LFD (lower is better) scores for the **ablation studies** of each key component for the focal frequency loss (FFL).

	PSNR↑	SSIM↑	LPIPS↓	FID↓	LFD↓
baseline	20.044	0.568	0.237	97.035	14.785
full FFL	**21.703**	**0.642**	**0.199**	**83.801**	**14.403**
w/o freq	18.200	0.470	0.265	123.833	15.210
w/o phase	13.273	0.380	0.407	233.170	16.344
w/o ampli	15.640	0.359	0.539	323.528	15.799
w/o focal	20.163	0.574	0.234	94.497	14.758

5. Conclusion

The proposed focal frequency loss directly optimizes image reconstruction and synthesis methods in the frequency domain. The loss adaptively focuses the model on the frequency components that are hard to deal with to ameliorate quality. The loss is complementary to existing spatial losses of diverse baselines varying in categories, network structures, and tasks, outperforming relevant approaches. We further show the potential of focal frequency loss to improve synthesis results of StyleGAN2. Exploring other applications and devising better frequency domain optimization strategies can be interesting future works.

Acknowledgments. This study is supported under the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s).

References

[1] DeepFaceLab. https://github.com/iperov/DeepFaceLab/. Accessed: 2019-08-20. 2

[2] DeepFakes. https://github.com/deepfakes/faceswap/. Accessed: 2019-08-16. 2

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *ICLR*, 2018. 5

[4] Mu Cai, Hong Zhang, Huijuan Huang, Qichuan Geng, and Gao Huang. Frequency domain image translation: More photo-realistic, better identity-preserving. *arXiv preprint*, arXiv:2011.13611, 2020. 2

[5] Qifeng Chen and Vladlen Koltun. Photographic image synthesis with cascaded refinement networks. In *ICCV*, 2017. 7

[6] Wenlin Chen, James Wilson, Stephen Tyree, Kilian Q Weinberger, and Yixin Chen. Compressing convolutional neural networks in the frequency domain. In *KDD*, 2016. 2

[7] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*, 2018. 2

[8] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. Describing textures in the wild. In *CVPR*, 2014. 5

[9] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 5

[10] Ricard Durall, Margret Keuper, and Janis Keuper. Watch your up-convolution: CNN based generative deep neural networks are failing to reproduce spectral distributions. In *CVPR*, 2020. 2, 5, 8

[11] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *TPAMI*, 32:1627–1645, 2009. 2, 4

[12] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *ICML*, 2020. 2

[13] Manuel Fritsche, Shuhang Gu, and Radu Timofte. Frequency separation for real-world super-resolution. In *ICCVW*, 2019. 2

[14] Rinon Gal, Dana Cohen, Amit Bermano, and Daniel Cohen-Or. SWAGAN: A style-based wavelet-driven generative model. *arXiv preprint*, arXiv:2102.06108, 2021. 2

[15] 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. 1, 2

[16] Lionel Gueguen, Alex Sergeev, Ben Kadlec, Rosanne Liu, and Jason Yosinski. Faster neural networks straight from JPEG. In *NeurIPS*, 2018. 2

[17] Seungwook Han, Akash Srivastava, Cole Hurwitz, Prasanna Sattigeri, and David D Cox. not-so-biggan: Generating high-fidelity images on a small compute budget. *arXiv preprint*, arXiv:2009.04433, 2020. 2

[18] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017. 5

[19] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313:504–507, 2006. 1, 2, 5

[20] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. RDA: Robust domain adaptation via fourier adversarial attacking. In *ICCV*, 2021. 2

[21] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, 2018. 2

[22] Yihao Huang, Felix Juefei-Xu, Qing Guo, Xiaofei Xie, Lei Ma, Weikai Miao, Yang Liu, and Geguang Pu. FakeRetouch: Evading deepfakes detection via the guidance of deliberate noise. *arXiv preprint*, arXiv:2009.09213, 2020. 1, 2

[23] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 1, 2, 5, 6

[24] Liming Jiang, Zhengkui Guo, Wayne Wu, Zhaoyang Liu, Ziwei Liu, Chen Change Loy, et al. DeeperForensics Challenge 2020 on real-world face forgery detection: Methods and results. *arXiv preprint*, arXiv:2102.09471, 2021. 2

[25] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. DeeperForensics-1.0: A large-scale dataset for real-world face forgery detection. In *CVPR*, 2020. 2

[26] Liming Jiang, Changxu Zhang, Mingyang Huang, Chunxiao Liu, Jianping Shi, and Chen Change Loy. TSIT: A simple and versatile framework for image-to-image translation. In *ECCV*, 2020. 2, 5

[27] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. 2, 5, 8

[28] Steffen Jung and Margret Keuper. Spectral distribution aware image generation. *arXiv preprint*, arXiv:2012.03110, 2020. 2

[29] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. *arXiv preprint*, arXiv:1710.10196, 2017. 2, 5

[30] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 2, 7

[31] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *CVPR*, 2020. 1, 2, 5, 7

[32] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *NeurIPS*, 2018. 1

[33] Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In *NeurIPS*, 2014. 2

[34] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint*, arXiv:1312.6114, 2013. 1, 2, 5

[35] A Levinskis. Convolutional neural network feature reduction using wavelet transform. *Elektronika ir Elektrotechnika*, 19:61–64, 2013. 2

[36] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017. 2, 4

[37] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *NeurIPS*, 2017. 2

[38] Xihui Liu, Guojun Yin, Jing Shao, Xiaogang Wang, and Hongsheng Li. Learning to predict layout-to-image conditional convolutions for semantic image synthesis. In *NeurIPS*, 2019. 2

[39] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 5

[40] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 1, 2, 5

[41] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint*, arXiv:1411.1784, 2014. 2

[42] Augustus Odena, Vincent Dumoulin, and Chris Olah. Deconvolution and checkerboard artifacts. *Distill*, 1:e3, 2016. 1

[43] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, 2019. 1, 2, 5, 7

[44] Stanislav Pidhorskyi, Donald A Adjeroh, and Gianfranco Doretto. Adversarial latent autoencoders. In *CVPR*, 2020. 1

[45] Xiaojuan Qi, Qifeng Chen, Jiaya Jia, and Vladlen Koltun. Semi-parametric image synthesis. In *CVPR*, 2018. 7

[46] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint*, arXiv:1511.06434, 2015. 2

[47] Radim Šára Radim Tyleček. Spatial pattern templates for recognition of objects with regular structure. In *GCPR*, 2013. 5

[48] 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 *ICML*, 2019. 1, 2

[49] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *NeurIPS*, 2008. 2

[50] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *NeurIPS*, 2016. 5

[51] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *CVPR*, 2016. 2, 4

[52] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint*, arXiv:1409.1556, 2014. 8

[53] Nalini M Singh, Juan Eugenio Iglesias, Elfar Adalsteinsson, Adrian V Dalca, and Polina Golland. Joint frequency-and image-space learning for fourier imaging. *arXiv preprint*, arXiv:2007.01441, 2020. 2

[54] Matthew Tancik, Pratul P Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In *NeurIPS*, 2020. 1, 2

[55] Arash Vahdat and Jan Kautz. NVAE: A deep hierarchical variational autoencoder. In *NeurIPS*, 2020. 1

[56] Aaron Van den Oord, Nal Kalchbrenner, Lasse Espeholt, Oriol Vinyals, Alex Graves, et al. Conditional image generation with PixelCNN decoders. In *NeurIPS*, 2016. 1

[57] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 2

[58] Haohan Wang, Xindi Wu, Zeyi Huang, and Eric P Xing. High-frequency component helps explain the generalization of convolutional neural networks. In *CVPR*, 2020. 2

[59] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. CNN-generated images are surprisingly easy to spot...for now. In *CVPR*, 2020. 1, 2

[60] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional GANs. In *CVPR*, 2018. 2, 7

[61] Yunhe Wang, Chang Xu, Shan You, Dacheng Tao, and Chao Xu. CNNpack: Packing convolutional neural networks in the frequency domain. In *NeurIPS*, 2016. 2

[62] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *TIP*, 13:600–612, 2004. 5

[63] Yunxuan Wei, Shuhang Gu, Yawei Li, and Longcun Jin. Unsupervised real-world image super resolution via domain-distance aware training. *arXiv preprint*, arXiv:2004.01178, 2020. 2

[64] Mingqing Xiao, Shuxin Zheng, Chang Liu, Yaolong Wang, Di He, Guolin Ke, Jiang Bian, Zhouchen Lin, and Tie-Yan Liu. Invertible image rescaling. In *ECCV*, 2020. 2

[65] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *ECCV*, 2018. 5

[66] Kai Xu, Minghai Qin, Fei Sun, Yuhao Wang, Yen-Kuang Chen, and Fengbo Ren. Learning in the frequency domain. In *CVPR*, 2020. 2

[67] Zhi-Qin John Xu, Yaoyu Zhang, Tao Luo, Yanyang Xiao, and Zheng Ma. Frequency principle: Fourier analysis sheds light on deep neural networks. *arXiv preprint*, arXiv:1901.06523, 2019. 1, 2

[68] Jaejun Yoo, Youngjung Uh, Sanghyuk Chun, Byeongkyu Kang, and Jung-Woo Ha. Photorealistic style transfer via wavelet transforms. In *ICCV*, 2019. 2

[69] Aron Yu and Kristen Grauman. Fine-grained visual comparisons with local learning. In *CVPR*, 2014. 5

[70] Fisher Yu, Vladlen Koltun, and Thomas Funkhouser. Dilated residual networks. In *CVPR*, 2017. 5

[71] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 5

[72] Xu Zhang, Svebor Karaman, and Shih-Fu Chang. Detecting and simulating artifacts in gan fake images. In *WIFS*, 2019. 1, 2

[73] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In *CVPR*, 2017. 5

[74] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 2