This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Frequency Domain Image Translation: More Photo-realistic, Better Identity-preserving

Mu Cai[1] Hong Zhang[2] Huijuan Huang[3] Qichuan Geng[4] Yixuan Li[1] Gao Huang[5]

[1]University of Wisconsin-Madison [2]SenseTime Group Ltd. [3]Kwai Inc.
[4]Beihang University [5] Tsinghua University

{mucai,sharonli}@cs.wisc.edu fykalviny@gmail.com huanghuijuan@kuaishou.com
zhaokefirst@buaa.edu.cn gaohuang@tsinghua.edu.cn

Abstract

Image-to-image translation has been revolutionized with GAN-based methods. However, existing methods lack the ability to preserve the identity of the source domain. As a result, synthesized images can often over-adapt to the reference domain, losing important structural characteristics and suffering from suboptimal visual quality. To solve these challenges, we propose a novel frequency domain image translation (FDIT) framework, exploiting frequency information for enhancing the image generation process. Our key idea is to decompose the image into low-frequency and high-frequency components, where the high-frequency feature captures object structure akin to the identity. Our training objective facilitates the preservation of frequency information in both pixel space and Fourier spectral space. We broadly evaluate FDIT across five large-scale datasets and multiple tasks including image translation and GAN inversion. Extensive experiments and ablations show that FDIT effectively preserves the identity of the source image, and produces photo-realistic images. FDIT establishes ***state-of-the-art*** *performance, reducing the average FID score by 5.6% compared to the previous best method.*

1. Introduction

Image-to-image translation [67, 9, 4, 56, 53] has attracted great research attention in computer vision, which is tasked to synthesize new images based on the source and reference images (see Figure 1). This task has been revolutionized since the introduction of GAN-based methods [28, 66]. In particular, a plethora of literature attempts to decompose the image representation into a content space and a style space [11, 45, 37, 26]. To translate a source image, its content representation is combined with a different style representation from the reference domain.

Despite exciting progress, existing solutions suffer from two notable challenges. First, there is no explicit mechanism that allows preserving the identity, and as a result, the synthesized image can over-adapt to the reference domain and lose the original identity characteristics. This can be observed in Figure 1, where Swapping Autoencoder [45]

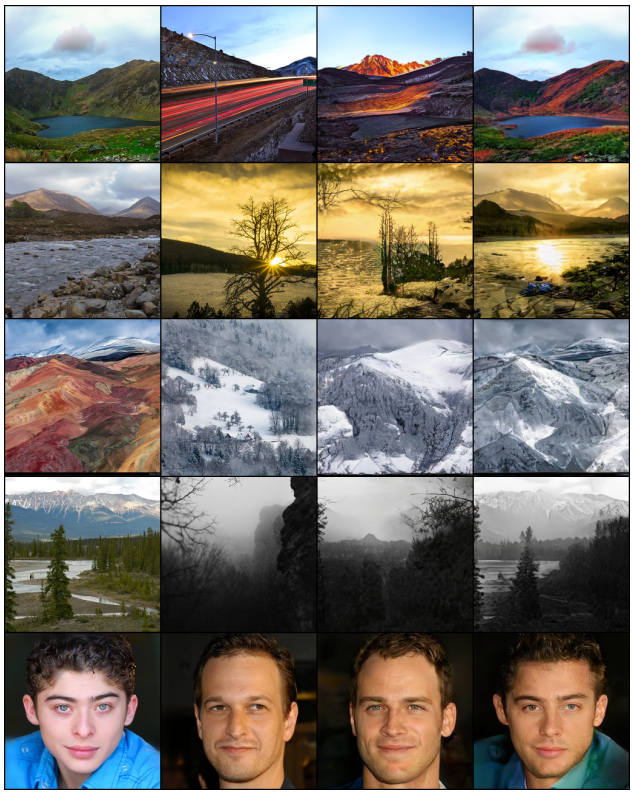

Figure 1: Image translation results of the Flicker mountains dataset. From left column to right: we show the source images, reference images, the generated images using Swapping Autoencoder [45] and FDIT (ours), respectively. SwapAE over-adapt to the reference image. FDIT better preserves the composition and identity with respect to the source image.

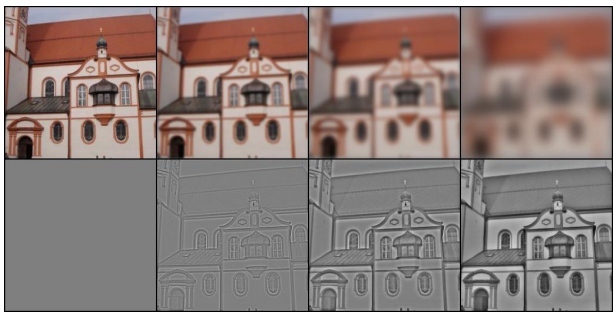

Figure 2: Visualization of the effect of decomposing the original image into grayscale high frequency (*bottom*) and low frequency (*top*) components. Gaussian kernel is employed as the low-frequency filter with different kernel sizes k.

generates images with identity and structure closer to the reference rather than the source image. For example, in the second row, the tree is absent from the source image yet occurs in the translation result. Second, the generation process may lose important fine-grained details, leading to suboptimal visual quality. This can be prohibitive for generating photo-realistic high-resolution images. The challenges above raise the following important question: *how can we enable photo-realistic image translation while better preserving the identity?*

Motivated by this, we propose a novel framework–*Frequency Domain Image Translation* (**FDIT**)–exploiting frequency information for enhancing the image generation process. Our key idea is to decompose the image into low- and high-frequency components, and regulate the frequency consistency during image translation. Our framework is inspired by and grounded in signal processing [15, 5, 21]. Intuitively, the low-frequency component captures information such as color and illumination; whereas the high-frequency component corresponds to sharp edges and important details of objects. For example, Figure 2 shows the resulting images via adopting the Gaussian blur to decompose the original image into low- *vs*. high-frequency counterparts (top *vs*. bottom). The building identity is distinguishable based on the high-frequency components.

Formally, FDIT introduces novel frequency-based training objectives, which facilitate the preservation of frequency information during training. The frequency information can be reflected in the visual space as identity characteristics and important fine details. Formally, we impose restrictions in both *pixel space* as well as the *Fourier spectral space*. In the pixel space, we transform each image into its high-frequency and low-frequency components by applying the Gaussian kernel (*i.e*., low-frequency filter). A loss term regulates the high-frequency components to be similar between the source image and the generated image. Furthermore, FDIT directly regulates the consistency in the frequency domain by applying Fast Fourier Transformation (FFT) to each image. This additionally ensures that the original and translated images share a similar high-frequency spectrum.

Extensive experiments demonstrate that FDIT is highly effective, establishing **state-of-the-art** performance on image translation tasks. Below we summarize our key results and contributions:

- We propose a novel frequency-based image translation framework, FDIT, which substantially improves the identity-preserving generation, while enhancing the image hybrids realism. FDIT outperforms competitive baselines by a large margin, across all datasets considered. Compared to the vanilla Swapping Autoencoder (SwapAE) [45], FDIT decreases the FID score by **5.6%**.
- We conduct extensive ablations and user study to evaluate the (1) identity-preserving capability and (2) image quality, where FDIT constantly surpasses previous methods. For example, the user study shows an average preference of **75.40**% and **64.39**% for FDIT over Swap AE in the above two aspects. We also conduct the ablation study to understand the efficacy of different loss terms and frequency supervision modules.
- We broadly evaluate our approach across five large-scale datasets (including two newly collected ones). Quantitative and qualitative evaluations on image translation and GAN-inversion tasks demonstrate the superiority of our method[1].

2. Background: Image-to-image Translation

Image-to-image translation aims at directly generating the synthesized image given a source image and an accompanying reference image. Existing algorithms commonly employ an encoder-decoder–like neural network architecture. We denote the encoder $E(\mathbf{x})$, the generator $G(\mathbf{z})$, and the image space $\mathcal{X} = \mathbb{R}^{H \times W \times 3}$ (RGB color channels).

Given an image $\mathbf{x} \in \mathcal{X}$, the encoder E maps it to a latent representation $\mathbf{z} \in \mathcal{Z}$. Previous approaches rely on the assumption that the latent code can be composed into two components $\mathbf{z} = (\mathbf{z}_c, \mathbf{z}_s)$, where $\mathbf{z}_c$ and $\mathbf{z}_s$ correspond to the **c**ontent and **s**tyle information respectively. A reconstruction loss minimizes the L_1 norm between the original input $\mathbf{x}$ and $G(E(\mathbf{x}))$.

To perform image translation, the generator takes the content code $\mathbf{z}_c^{\text{source}}$ from the source image, together with the style code $\mathbf{z}_s^{\text{ref}}$ from the reference image. The translated

[1]Code and dataset are available at: https://github.com/mu-cai/frequency-domain-image-translation

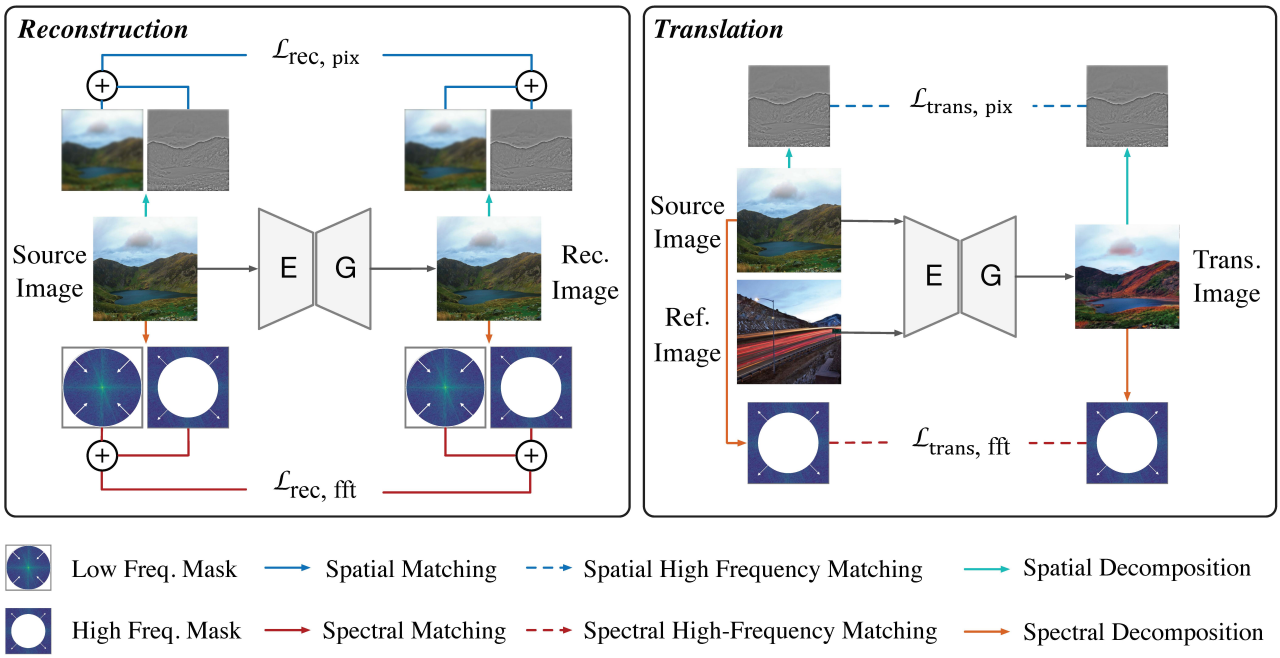

Figure 3: Overview of the proposed *frequency domain image translation* (**FDIT**) framework. The key idea is to decompose the image into low-frequency and high-frequency components, and regulate the frequency consistency during image reconstruction (*left*) and image translation (*right*). High frequency information captures the sharp edges and important details of objects, where is effectively matched by FDIT training objectives.

image is given by $G(\mathbf{z}_c^{\text{source}}, \mathbf{z}_s^{\text{ref}})$. However, existing methods can be limited by its feature disentanglement ability, where $\mathbf{z}_c^{\text{source}}$ may not capture the identity of source image. As a result, such identity-related characteristics can be undesirably *lost in translation* (see Figure 5), which motivates our work.

3. Frequency Domain Image Translation

Our novel frequency-based image translation framework is illustrated in Figure 3. In what follows, we first provide an overview and then describe the training objective. Our training objective facilitates the preservation of frequency information during the image translation process. Specifically, we impose restrictions in both *pixel space* (Section 3.1) as well as the *Fourier spectral space* (Section 3.2).

3.1. Pixel Space Loss

High- and low-frequency images. We transform each input $\mathbf{x}$ into two images $\mathbf{x}_L \in \mathcal{X}$ and $\mathbf{x}_H \in \mathcal{X}$, which correspond to the low-frequency and high-frequency images respectively. Note that both $\mathbf{x}_L$ and $\mathbf{x}_H$ are in the same spatial dimension as $\mathbf{x}$. Specifically, we employ the Gaussian kernel, which filters the high frequency feature and keeps the low frequency information:

$$k_\sigma[i,j] = \frac{1}{2\pi\sigma^2} e^{-\frac{1}{2}\left(\frac{i^2+j^2}{\sigma^2}\right)}, \tag{1}$$

where $[i,j]$ denotes the spatial location within the image, and σ^2 denotes the variance of the Gaussian function. Following [21], the variance is increased proportionally with the Gaussian kernel size . Using convolution of the Gaussian kernel on input $\mathbf{x}$, we obtain the low frequency (*blurred*) image $\mathbf{x}_L$:

$$\mathbf{x}_L[i,j] = \sum_m \sum_n k[m,n] \cdot \mathbf{x}[i+m, j+n]. \tag{2}$$

where m, n denotes the index of an 2D Gaussian kernel, *i.e.*, $m, n \in [-\frac{k-1}{2}, \frac{k-1}{2}]$.

To obtain $\mathbf{x}_H$, we first convert color images into grayscale, and then subtract the low frequency information:

$$\mathbf{x}_H = \texttt{rgb2gray}(\mathbf{x}) - (\texttt{rgb2gray}(\mathbf{x}))_L, \tag{3}$$

where the `rgb2gray` function converts the color image to the grayscale. This removes the color and illumination information that is unrelated to the identity and structure. The resulting high frequency image $\mathbf{x}_H$ contains the sharp edges, *i.e. sketch* of the original image.

Reconstruction loss in the pixel space. We now employ the following reconstruction loss term, which enforces the similarity between the input and generator's output, for both

low-frequency and high-frequency components:

$$\mathcal{L}_{\text{rec,pix}}(E,G) = \mathbb{E}_{\mathbf{x}\sim\mathcal{X}}\Big[\|\mathbf{x}_L - (G(E(\mathbf{x})))_L\|_1 + \|\mathbf{x}_H - (G(E(\mathbf{x})))_H\|_1\Big]. \tag{4}$$

Translation matching loss in the pixel space. In addition to reconstruction loss, we also employ the translation matching loss:

$$\mathcal{L}_{\text{trans,pix}}(E,G) = \mathbb{E}_{\mathbf{x}\sim\mathcal{X}}\left[\|\mathbf{x}_H^{\text{source}} - \left(G(\mathbf{z}_c^{\text{source}}, \mathbf{z}_s^{\text{ref}})\right)_H\|_1\right], \tag{5}$$

where $\mathbf{z}_c^{\text{source}}$ and $\mathbf{z}_s^{\text{ref}}$ are the content code of the source image and the style code of the reference image, respectively. Intuitively, the translated images should adhere to the identity of the original image. We achieve this by regulating the high frequency components, and enforce the generated image to have the same high frequency images as the original source image.

3.2. Fourier Frequency Space Loss

Transformation from pixel space to the Fourier spectral space. In addition to the pixel-space constraints, we introduce loss terms that directly operate in the Fourier domain space. In particular, we use Fast Fourier Transformation (FFT) and map $\mathbf{x}$ from the pixel space to the Fourier spectral space. We apply the Discrete Fourier Transform $\mathcal{F}$ on a real 2D image I of size $H \times W$:

$$\mathcal{F}(I)(a,b) = \frac{1}{HW}\sum_{h=0}^{H-1}\sum_{w=0}^{W-1} e^{-2\pi i\cdot\frac{ha}{H}} e^{-2\pi i\cdot\frac{wb}{W}}\cdot I(h,w), \tag{6}$$

for $a = 0,\ldots,H-1, b = 0,\ldots,W-1$.

For the ease of post processing, we then transform $\mathcal{F}$ from the complex number domain to the real number domain. Additionally, we take the logarithm to stabilize the training:

$$\mathcal{F}^R(I)(a,b) = \log(1 + \sqrt{[\text{Re}\mathcal{F}(I)(a,b)]^2} + \sqrt{[\text{Im}\mathcal{F}(I)(a,b)]^2 + \epsilon}), \tag{7}$$

where $\epsilon = 1\times10^{-8}$ is a term added for numerical stability; Re and Im denote the real part and the imaginary part of $\mathcal{F}(I)(a,b)$ respectively. Each point in the Fourier spectrum would utilize information from all pixels according to the discrete spatial frequency, which would represent the frequency features in the global level.

Reconstruction loss in the Fourier space We then regulate the reconstruction loss in the frequency spectrum:

$$\mathcal{L}_{\text{rec,fft}}(E,G) = \mathbb{E}_{\mathbf{x}\sim\mathcal{X}}\left[\|\mathcal{F}^R(\mathbf{x}) - \mathcal{F}^R(G(E(\mathbf{x})))\|_1\right]. \tag{8}$$

Translation matching loss in the Fourier space. In a similar spirit as Equation 5, we devise a translation matching loss in the Fourier frequency domain:

$$\mathcal{L}_{\text{trans,fft}}(E,G) = \mathbb{E}_{\mathbf{x}\sim\mathcal{X}}\left[\|\mathcal{F}_H^R(\mathbf{x}^{\text{source}}) - \mathcal{F}_H^R(G\left(\mathbf{z}_c^{\text{source}}, \mathbf{z}_s^{\text{ref}}\right))\|_1\right], \tag{9}$$

where $\mathcal{F}_H^R(\mathbf{x}) = \mathcal{F}^R(\texttt{rgb2gray}(\mathbf{x}))\cdot M_H$. M_H is the frequency mask, for which we provided detailed explanation below. The loss constrains the high frequency components of the generated images for better identity preserving.

Frequency mask. As illustrated in Figure 3, the low-frequency mask is a circle with radius r, whereas the high-frequency mask is the complement region. The frequency masks M_H and M_L can be estimated empirically from the distribution of $\mathcal{F}^R$ on the entire training dataset. We choose the radius to be 21 for images with resolution 256×256. The energy within the low-frequency mask accounts for 97.8% of the total energy in the spectrum.

3.3. Overall Loss

Considering all the aforementioned losses, the overall loss is formalized as:

$$\mathcal{L}_{\text{FDIT}} = \mathcal{L}_{\text{org}} + \lambda_1\mathcal{L}_{\text{rec,pix}} + \lambda_2\mathcal{L}_{\text{trans,pix}} + \lambda_3\mathcal{L}_{\text{rec,fft}} + \lambda_4\mathcal{L}_{\text{trans,fft}}, \tag{10}$$

where $\mathcal{L}_{\text{org}}$ is the original loss function of *any* image translation model. For simplicity, we use $\lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = 1$ in this paper.

Gaussian kernel *vs*. FFT. Gaussian kernel and FFT are complementary for preserving the frequency information. On one hand, the Gaussian kernel extracts the frequency information via the convolution, therefore representing the frequency features in a *local* manner. On the other hand, Fast Fourier Transformation utilizes the information from all pixels to obtain the FFT value for each spatial frequency, characterizing the frequency distribution *globally*. Gaussian kernel and FFT are therefore complementary in preserving the frequency information. We show ablation study on this in Section 4.2, where both are effective in enhancing the identity-preserving capability for image translation tasks.

Gaussian kernel size When transforming the images in Figure 2 into the spectrum space, the effects of the Gaussian kernel size could be clearly reflected in Figure 4. To be specific, a large kernel would cause severe distortion on the low-frequency band while a small kernel would not preserve much of the high-frequency information. In this work, we choose the kernel size $k = 21$ for images with resolution 256×256, which could appropriately separate the high/low-frequency information, demonstrated in both image space

and spectral space distribution. Our experiments also show that FDIT is not sensitive to the selection of k as long as it falls into a mild range.

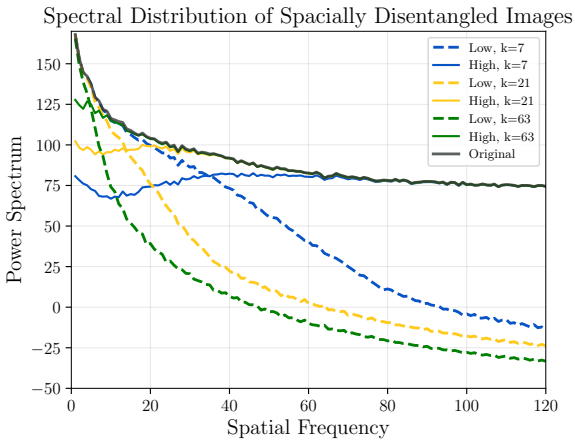

Figure 4: Transforming the resulting high- and low-frequency images in Figure 2 into the frequency power spectrum. The Gaussian kernel with kernel size $k = 21$ could avoid the distortion in high-frequency and low-frequency regions. The power spectrum represents the energy distribution at each spatial frequency.

4. Experiments

In this section, we evaluate our proposed method on two state-of-the-art image translation architectures, *i.e.*, Swapping Autoencoder [45], StarGAN v2 [11], and one GAN inversion model, *i.e.*, Image2StyleGAN [1]. Extensive experimental results show that FDIT not only better preserves the identity, but also enhances image quality.

Datasets. We evaluate FDIT on the following five datasets: (1) LSUN Church [62], (2) CelebA-HQ [32], (3) LSUN Bedroom [62], (4) Flickr Mountains (100k self-collected images), (5) Flickr Waterfalls (100k self-collected images). (6) Flickr Faces HQ (FFHQ) dataset [33]. All the images are trained and tested at 256×256 resolution except FFHQ, which is trained at 512×512, and finetuned at 1024×1024 resolution. For evaluation, we use a validation set that is separate from the training data.

4.1. Autoencoder

Autoencoder is widely used as the backbone of the deep image translation task [1, 26]. We use state-of-the-art Swapping Autoencoder (SwapAE) [45], which is built on the backbone of StyleGAN2 [34]. Swap AE also uses the technique in PatchGAN [29] to further improve the texture transferring performance. We incorporate our proposed FDIT training objectives into the vanilla SwapAE.

FDIT better preserves the identity with respect to the source image. We contrast the image translation performance using FDIT *vs.* vanilla SwapAE in Figure 1 and Figure 5. The vanilla SwapAE is unable to preserve the important identity of the source images, and over-adapts to the reference image. For example, the face identity is completely switched after translation, as seen in rows 4 of Figure 5. SwapAE also fails to preserve the outline and the local sharp edges in the source image. As shown in Figure 1, the outlines of the mountains are severely distorted. Besides, the overall image composition has a large shift from the original source image. In contrast, using our method FDIT, the identity and structure of the swapped hybrid images are highly preserved. As shown in Figure 1 and Figure 5, the overall sketches and local fine details are well preserved while the coloring, illumination, and even the weather are well transferred from the reference image (top rows of Figure 1).

Lastly, we compare FDIT with the state-of-the-art image stylization method STROTSS [35] and WCT2 [60]. Image stylization is a strong baseline as it emphasizes on the strict adherence to the source image. However, as shown in Figure 5, WCT2 leads to poor transferability in image generation tasks. Despite strong identity-preservation, STROTSS and WCT2 are less flexible, and generate images that highly resemble the source image. In contrast, FDIT can both preserve the identity of the source image as well as maintain a high transfer capability. This further demonstrates the superiority of FDIT in image translation.

FDIT enhances the image generation quality. We show in Table 1 that FDIT can substantially improve the image quality while preserving the image content. We adopt the Fréchet Inception Distance (FID) [22] as the measure of image quality. Small values indicate better image quality. Details about Im2StyleGAN [1] and StyleGAN2 [1] are shown in the supplementary material. FDIT achieves the lowest FID across all datasets. On average, FDIT could reduce the FID score by **5.6**% compared to the current state-of-the-art method.

Method \ Dataset	Church	Waterfalls	FFHQ	CelebA-HQ
Im2StyleGAN [1]	219.50	267.25	123.13	-
StyleGAN2 [1]	57.54	57.46	81.44	-
Swap AE [45]	52.34	50.90	59.83	43.47
FDIT (ours)	**48.21**	**48.76**	**55.96**	**42.02**

Table 1: Comparison of FID score on four diverse datasets: LSUN Church, Waterfalls, FFHQ and CelebA-HQ.

FDIT enables continuous interpolation between different domains. We show that FDIT enables image attribute editing task, which creates a series of smoothly changing images between two sets of distinct images [45, 48]. Our method performs image editing towards the target domain while strictly adhering to the content of the source image. We also verify the disentangled semantic latent vectors

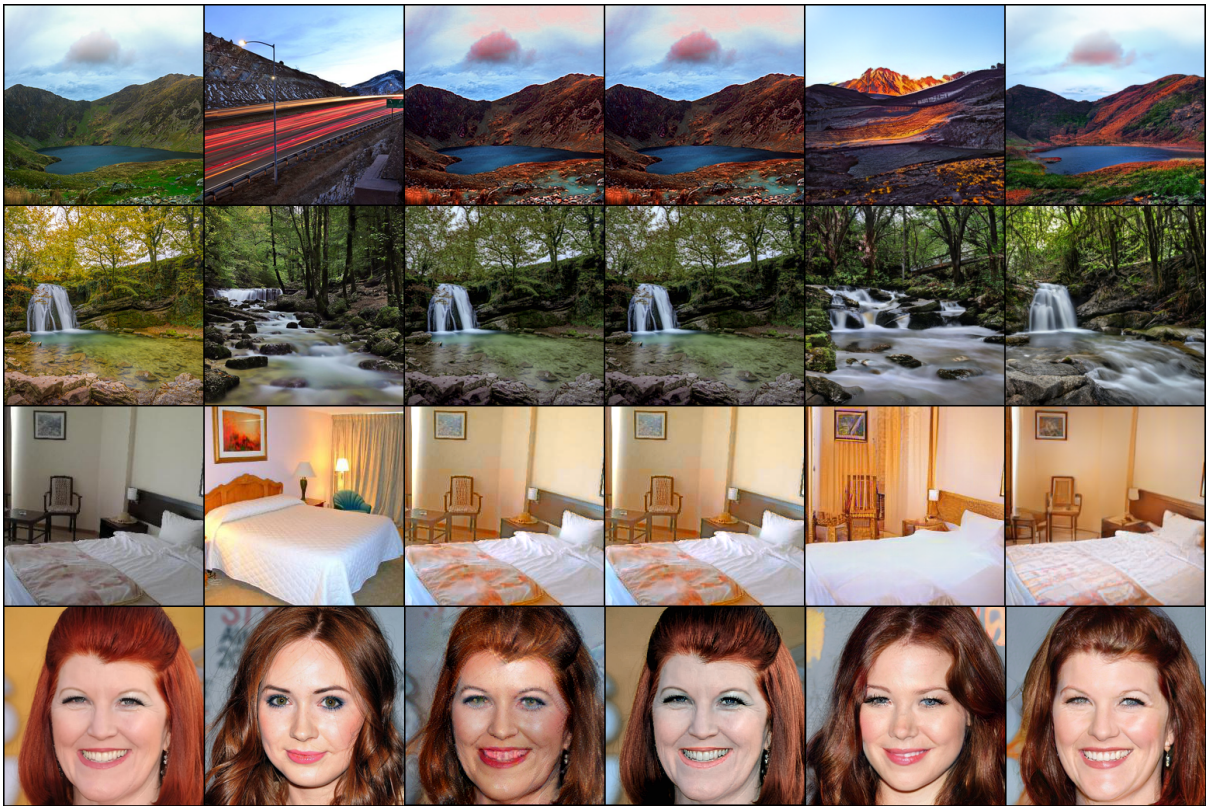

Figure 5: Results across four diverse datasets, including Flicker Mountains, Flicker Waterfalls, LSUN Bedroom [62] , and CelebA-HQ [32]. Swap AE [45] over-adapts to the reference image after image translation. In contrast, **FDIT** (ours) can better preserve identity of the source image. Compared to STROTSS [35] and WCT2 [60], FDIT can synthesize photo-realistic images. Zoom in for details.

using Principal Component Analysis (PCA). The identity-preserving results are shown in the supplementary material.

4.2. Ablation Study

We conduct ablation experiments of the frequency domain supervision in local and global levels on LSUN church dataset and compare them with the baseline Swapping Autoencoder [45] in terms of FID. As shown in Table. 2, we can find that both local and global manner can boost the baseline, and the final performance could be more boosted by adding both of them.

Pixel and Fourier space losses are complementary. To better understand our method, we isolate the effect of *pixel space loss* and *Fourier spectral space loss*. The results on the LSUN Church dataset are summarized in Table 2. The vanilla SwapAE is equivalent to having neither loss terms, which yields the FID score of 52.34. Using pixel space frequency loss reduces the FID score to 49.47. Our method is most effective when combining both pixel-space and Fourier-space loss terms, achieving the FID score of 48.21. Our ablation signifies the importance of using frequency-based training objectives.

Loss terms		
Pixel space	Fourier space	FID ↓
✗	✗	52.34
✓	✗	49.47
✗	✓	49.62
✓	✓	**48.21**

Table 2: Ablation study on the effect of pixel space loss and Fourier spectral space loss. Evaluations are based on the LSUN Church dataset.

4.3. GAN Inversion

FDIT improves reconstruction quality in GAN inversion. We evaluate the efficacy of FDIT on the GAN inversion task, which maps the real images into the noise latent vectors. In particular, Image2StyleGAN[1] serves as a strong baseline, which performs reconstruction between the real image and the generated images via iterative optimization over the latent vector.

We adopt the same architecture, however impose our frequency-based reconstruction loss. The inversion results are shown in Figure 6. On high-resolution (1024×1024) images, the quality of the inverted images is improved

Metrics \ Method	Image2StyleGAN	FDIT
MSE ↓	0.0226	**0.0205**
MAE ↓	0.0969	**0.0860**
PSNR ↑	19.626	**20.466**
SSIM ↑	0.6160	**0.6218**

Table 3: GAN inversion performance comparison, measured by the image reconstruction quality between Image2StyleGAN and FDIT (ours). Evaluation metrics includes mean-square error (MSE), mean absolute error (MAE), peak signal-to-noise ratio (PSNR) [14], and SSIM [55]. ↑ means that higher value represents better image quality, and vice versa.

across all scenes. FDIT better preserves the overall structure, fine details, and color distribution. We further measure the performance quantitatively, summarizing the results in Table 3. Under different metrics (MSE, MAE, PSNR, SSIM), our method FDIT outperforms Image2StyleGAN.

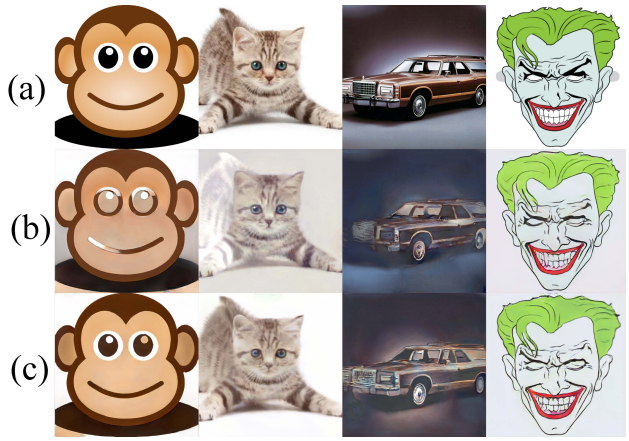

Figure 6: GAN inversion results on high resolution images (1024×1024). We compare (a) high-resolution source images, (b) Image2StyleGAN[1] results and (c) inverted images by FDIT (ours). FDIT better maintains fine details and visual quality.

4.4. StarGAN v2

StarGAN v2 is another state-of-the-art image translation model which can generate image hybrids guided by either reference images or latent noises. Similar to the autoencoder-based network, we can optimize the StarGAN v2 framework with our frequency-based losses. In order to validate FDIT in a stricter condition, we construct a CelebA-HQ-Smile dataset based on the smiling attribute from CelebA-HQ dataset. The style refers to whether that person smiles, and the content refers to the identity.

Several salient observations can be drawn from Figure 7. First, FDIT can highly preserve the gender identity; whereas the vanilla StarGAN v2 model would change the resulting gender according to the reference image (*e.g.* first and second row). Secondly, the image quality of FDIT is better, where FID is improved from 17.32 to 16.86. Thirdly, our model can change the smiling attribute while maintaining other facial features strictly. For example, as shown in the third row, StarGAN v2 undesirably changes the hairstyle from straight (source) to curly (reference), whereas FDIT maintains the same hairstyle.

Figure 7: Compared to vanilla StarGAN v2 [11], FDIT achieves much better identity-preserving ability.

4.5. User Study

We conduct a user study to qualitatively measure the generated images. Specifically, we employ the two-alternative forced-choice setting, which was commonly used to train Learned Perceptual Image Patch Similarity (LPIPS) [64] and to evaluate style transfer methods. We provide users with the source image, reference image, images generated by FDIT, and the baseline SOTA models. Each user is forced to choose which of the two image hybrids 1) better preserves the identity characteristics, and 2) has better image quality. We collected a total of 2,058 user preferences across 5 diverse datasets. Results are summarized in Table 4. On average, **75.40%** of preferences are given to FDIT for identity preserving; and **64.39%** of answers indicate FDIT produces more photo-realistic images.

Furthermore, comparing to StarGAN v2, 57.14% user preferences are given to FDIT for better content preservation; 53.34% user preferences indicate that FDIT produces better image quality compared to Image2StyleGAN. Therefore, the user study also verifies that FDIT produces better identity-preserving and photo-realistic images.

5. Related work

Generative adversarial networks (GAN). GAN [19, 20, 3, 6, 63, 47] has revolutionized revolutionized many com-

Ratio(%) / Metric / Dataset	Identity Preserving	Image Realism
LSUN Church	63.27	57.14
LSUN Bedroom	71.43	78.57
Flicker Mountains	80.10	66.84
Flicker Waterfalls	80.61	62.24
CelebA-HQ	57.14	53.06
Average	**75.40**	**64.39**

Table 4: Results of the user study on five datasets, which shows the preference of FDIT over Swapping Autoencoder [45] w.r.t identity preserving and image quality.

puter vision tasks, such as super resolution [36, 52], colorization [27, 61], and image synthesis [7, 42, 16]. Early work [46, 25] directly used the Gaussian noises as inputs to the generator. However, such an approach has unsatisfactory performance in generating photo-realistic images. Recent works significantly improved the image reality by injecting the noises hierarchically [33, 34] in the generator. These works adopt the adaptive instance normalization (AdaIN) module [23] for image stylization.

Image-to-image translation. Image-to-image translation [67, 51] synthesizes images by following the style of a reference image while keeping the content of the source image. One way is to use the GAN inversion, which maps the input from the pixel space into the latent noises space via the optimization method [1, 2, 34]. However, these methods are known to be computationally slow due to their iterative optimization process, which makes deployment in mobile devices difficult [1]. Furthermore, the quality of the reconstructed images can be suboptimal. Another approach is to utilize the conditional GAN (or autoencoder) to convert the input images into latent vectors [26, 10, 11, 45, 44, 43], making the image translation process much faster than GAN inversion. However, exiting state-of-the-art image translation models such as StarGAN v2 [11] and Swapping Autoencoder [45] can lose important structural characteristics of the source image. In this paper, we show that frequency-based information can effectively preserve the identity of the source image and enhance photo-realism.

Frequency domain in deep learning. Frequency domain analysis is widely used in traditional image processing [21, 12, 49, 31, 18]. The key idea of frequency analysis is to map the pixels from the Euclidean space to a frequency space, based on the changing speed in the spatial domain. Several works tried to bridge the connection between deep learning and frequency analysis [57, 8, 58, 59, 54, 41]. Chen *et al*. [8] and Xu *et al*. [57] showed that by incorporating frequency transformation, the neural network could be more efficient and effective. Wang *et al*. [50] found that the high-frequency components are useful in explaining the generalization of neural networks. Recently, Durall *et al*. [17] observed that the images generated by GANs are heavily distorted in high-frequency parts, and they introduced a spectral regularization term to the loss function to alleviate this problem. Czolbe *et al*. [13] proposed a frequency-based reconstruction loss for VAE using discrete Fourier Transformation (DFT). However, this approach does not incorporate pixel space frequency information, and relies on a separate dataset to get its free parameters. In fact, no prior work has explored using frequency-domain analysis for the image-to-image translation task. In this work, we explicitly devise a novel frequency domain *image translation* framework and demonstrate its superiority in performance.

Neural style transfer. Neural style transfer aims at transferring the low-level styles while strictly maintaining the content in the source image [60, 35, 24, 39, 38, 40]. Typically, the texture is represented by the global image statistics while the content is controlled by the perception metric [60, 30, 65]. However, existing methods could only handle the local color transformation, making it hard to transform the overall style and semantics. More specifically, they struggle in the cross-domain image translations, for example, gender transformation [60]. In other words, despite strong identity-preservation ability, such methods are less flexible for the cross-domain translation and can generate images that highly resemble the source domain. In contrast, FDIT can both preserve the identity of the source images while maintaining a high domain transfer capability.

6. Conclusion

In this paper, we propose *Frequency Domain Image Translation (FDIT)*, a novel image translation framework that preserves the frequency information in both pixel space and Fourier spectral space. Unlike the existing image translation models, FDIT directly uses high-frequency components to capture object structure akin to the identity. Experimental results on five large-scale datasets and multiple tasks show that FDIT effectively preserves the identity of the source image while producing photo-realistic image hybrids. Extensive user study and ablations further validate the effectiveness of our approach both qualitatively and quantitatively. We hope future research will increase the attention towards frequency-based approaches for image translation tasks.

7. Acknowledgment

Mu Cai and Yixuan Li are supported by funding from the Wisconsin Alumni Research Foundation (WARF). Gao Huang is supported in part by the National Key R&D Program of China under Grant 2020AAA0105200, the National Natural Science Foundation of China under Grants 62022048 and 61906106, the Institute for Guo Qiang of Tsinghua University and Beijing Academy of Artificial Intelligence.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *IEEE International Conference on Computer Vision*, 2019. 5, 6, 7, 8

[2] R. Abdal, Y. Qin, and P. Wonka. Image2stylegan++: How to edit the embedded images? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 8

[3] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *Proceedings of the International Conference on Machine Learning*, 2017. 7

[4] Deblina Bhattacharjee, Seungryong Kim, Guillaume Vizier, and Mathieu Salzmann. Dunit: Detection-based unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 1

[5] E Oran Brigham. *The fast Fourier transform and its applications*. Prentice-Hall, Inc., 1988. 2

[6] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019. 7

[7] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019. 8

[8] Yunpeng Chen, Haoqi Fan, Bing Xu, Zhicheng Yan, Yannis Kalantidis, Marcus Rohrbach, Shuicheng Yan, and Jiashi Feng. Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution. In *IEEE International Conference on Computer Vision*, 2019. 8

[9] Ying-Cong Chen, Xiaogang Xu, and Jiaya Jia. Domain adaptive image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 1

[10] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. 8

[11] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 1, 5, 7, 8

[12] James W Cooley. The re-discovery of the fast fourier transform algorithm. *Microchimica Acta*, 93(1):33–45, 1987. 8

[13] Steffen Czolbe, Oswin Krause, Ingemar J. Codex, and Christian Igel. A loss function for generative neural networks based on watson's perceptual model. In *Advances in Neural Information Processing Systems*, 2020. 8

[14] Johannes F De Boer, Barry Cense, B Hyle Park, Mark C Pierce, Guillermo J Tearney, and Brett E Bouma. Improved signal-to-noise ratio in spectral-domain compared with time-domain optical coherence tomography. *Optics letters*, 28(21):2067–2069, 2003. 7

[15] Guang Deng and LW Cahill. An adaptive gaussian filter for noise reduction and edge detection. In *IEEE Conference Record Nuclear Science Symposium and Medical Imaging Conference*, pages 1615–1619. IEEE, 1993. 2

[16] Jeff Donahue and Karen Simonyan. Large scale adversarial representation learning. In *Advancesin Neural Information Processing Systems*, 2019. 8

[17] Ricard Durall, Margret Keuper, and Janis Keuper. Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 8

[18] W Morven Gentleman and Gordon Sande. Fast fourier transforms: for fun and profit. In *Proceedings of the November 7-10, 1966, fall joint computer conference*, pages 563–578, 1966. 8

[19] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. In *Advancesin Neural Information Processing Systems*, 2014. 7

[20] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, 2017. 7

[21] Michael Heideman, Don Johnson, and Charles Burrus. Gauss and the history of the fast fourier transform. *IEEE ASSP Magazine*, 1(4):14–21, 1984. 2, 3, 8

[22] 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 *Advancesin Neural Information Processing Systems*, 2017. 5

[23] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *IEEE International Conference on Computer Vision*, 2017. 8

[24] X. Huang and S. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *2017 IEEE International Conference on Computer Vision*, 2017. 8

[25] Xun Huang, Yixuan Li, Omid Poursaeed, John Hopcroft, and Serge Belongie. Stacked generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017. 8

[26] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision*, 2018. 1, 5, 8

[27] Kim Hyunsu, Jhoo Ho Young, Park Eunhyeok, and Yoo Sungjoo. Tag2pix: Line art colorization using text tag with secat and changing loss. In *IEEE International Conference on Computer Vision*, 2019. 8

[28] P. Isola, J. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017. 1

[29] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017. 5

[30] Yongcheng Jing, Yezhou Yang, Zunlei Feng, Jingwen Ye, Yizhou Yu, and Mingli Song. Neural style transfer: A review. *IEEE Transactions on Visualization and Computer Graphics*, 2019. 8

[31] Steven G Johnson and Matteo Frigo. A modified split-radix fft with fewer arithmetic operations. *IEEE Transactions on Signal Processing*, 55(1):111–119, 2006. 8

[32] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018. 5, 6

[33] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 5, 8

[34] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 5, 8

[35] Nicholas Kolkin, Jason Salavon, and Gregory Shakhnarovich. Style transfer by relaxed optimal transport and self-similarity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 5, 6, 8

[36] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017. 8

[37] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Kumar Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European Conference on Computer Vision*, 2018. 1

[38] Xueting Li, Sifei Liu, Jan Kautz, and Ming-Hsuan Yang. Learning linear transformations for fast arbitrary style transfer. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 8

[39] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. In *Advances in Neural Information Processing Systems*, 2017. 8

[40] Yijun Li, Ming-Yu Liu, Xueting Li, Ming-Hsuan Yang, and Jan Kautz. A closed-form solution to photorealistic image stylization. In *Proceedings of the European Conference on Computer Vision*, September 2018. 8

[41] Zhenhua Liu, Jizheng Xu, Xiulian Peng, and Ruiqin Xiong. Frequency-domain dynamic pruning for convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2018. 8

[42] Mario Lucic, Michael Tschannen, Marvin Ritter, Xiaohua Zhai, Olivier Bachem, and Sylvain Gelly. High-fidelity image generation with fewer labels. In *Proceedings of the International Conference on Machine Learning*, 2019. 8

[43] Taesung Park, Alexei A. Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for conditional image synthesis. In *Proceedings of the European Conference on Computer Vision*, 2020. 8

[44] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 8

[45] Taesung Park, Jun-Yan Zhu, Oliver Wang, Jingwan Lu, Eli Shechtman, Alexei A. Efros, and Richard Zhang. Swapping autoencoder for deep image manipulation. In *Advancesin Neural Information Processing Systems*, 2020. 1, 2, 5, 6, 8

[46] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *International Conference on Learning Representations*, 2016. 8

[47] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, 2016. 7

[48] Y. Shen, C. Yang, X. Tang, and B. Zhou. Interfacegan: Interpreting the disentangled face representation learned by gans. *IEEE TPAMI*, pages 1–1, 2020. 5

[49] Charles Van Loan. *Computational frameworks for the fast Fourier transform*. SIAM, 1992. 8

[50] Haohan Wang, Xindi Wu, Zeyi Huang, and Eric P Xing. High-frequency component helps explain the generalization of convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 8

[51] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 8

[52] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision*, 2018. 8

[53] Yi Wang, Ying-Cong Chen, Xiangyu Zhang, Jian Sun, and Jiaya Jia. Attentive normalization for conditional image generation. *arXiv preprint arXiv:2004.03828*, 2020. 1

[54] Yunhe Wang, Chang Xu, Shan You, Dacheng Tao, and Chao Xu. Cnnpack: Packing convolutional neural networks in the frequency domain. In *Advances in Neural Information Processing Systems*, 2016. 8

[55] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 7

[56] Wei Xiong, Yutong He, Yixuan Zhang, Wenhan Luo, Lin Ma, and Jiebo Luo. Fine-grained image-to-image transformation towards visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 1

[57] Kai Xu, Minghai Qin, Fei Sun, Yuhao Wang, Yen-Kuang Chen, and Fengbo Ren. Learning in the frequency domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 8

[58] Zhi-Qin John Xu. Frequency principle: Fourier analysis sheds light on deep neural networks. *Communications in Computational Physics*, 28(5):1746–1767, Jun 2020. 8

[59] Zhi-Qin John Xu, Yaoyu Zhang, and Yanyang Xiao. Training behavior of deep neural network in frequency domain. In Tom Gedeon, Kok Wai Wong, and Minho Lee, editors, *Neural Information Processing*, pages 264–274, Cham, 2019. Springer International Publishing. 8

[60] Jaejun Yoo, Youngjung Uh, Sanghyuk Chun, Byeongkyu Kang, and Jung-Woo Ha. Photorealistic style transfer via wavelet transforms. In *IEEE International Conference on Computer Vision*, 2019. 5, 6, 8

[61] Seungjoo Yoo, Hyojin Bahng, Sunghyo Chung, Junsoo Lee, Jaehyuk Chang, and Jaegul Choo. Coloring with limited data: Few-shot colorization via memory augmented networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 8

[62] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 5, 6

[63] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *2017 IEEE International Conference on Computer Vision (IEEE International Conference on Computer Vision)*, 2017. 7

[64] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. 7

[65] Yexun Zhang, Ya Zhang, and Wenbin Cai. Separating style and content for generalized style transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, 2018. 8

[66] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision*, 2017. 1

[67] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *Advances in Neural Information Processing Systems*, 2017. 1, 8