Frequency Domain Image Translation: More Photo-realistic, Better Identity-preserving (Supplementary Material)

A. Image Attributes Editing Results

We demonstrate the identity preserving capability and photo realism of FDIT under the image attribute editing task via continuous interpolation and unsupervised semantic vector discovery.

A.1. Continuous interpolation between Different Domains

We show that FDIT can generate a series of smoothly changing images between two sets of distinct images. Vector arithmetic is one commonly used way to achieve this [10]. For example, we can sample n images from each of the two target domains, and then compute the average difference of the vectors between these two sets of images:

$$\hat{\mathbf{z}} = \frac{1}{n}\sum_{i=0}^{n} \mathbf{z}_i^{\mathbf{d1}} - \frac{1}{n}\sum_{j=0}^{n} \mathbf{z}_j^{\mathbf{d2}}, \tag{1}$$

where $\mathbf{z}^{\mathbf{d1}}, \mathbf{z}^{\mathbf{d2}}$ denote the latent code from two domains.

We perform interpolation on the style code while keeping the content code unchanged. The generated images can be formalized as $\mathbf{x}_{\text{gen}} = G(\mathbf{z}^{\text{source}}, \mathbf{z}^{\text{ref}} + \theta \cdot \hat{\mathbf{z}})$, where θ is the interpolation parameter. Figure 1 shows season transformation results using the Flicker Mountains dataset. Our identity-preserving image hybrids demonstrate that FDIT could achieve high-quality image editing performance towards the target domain while strictly adhering to the identity of the source image.

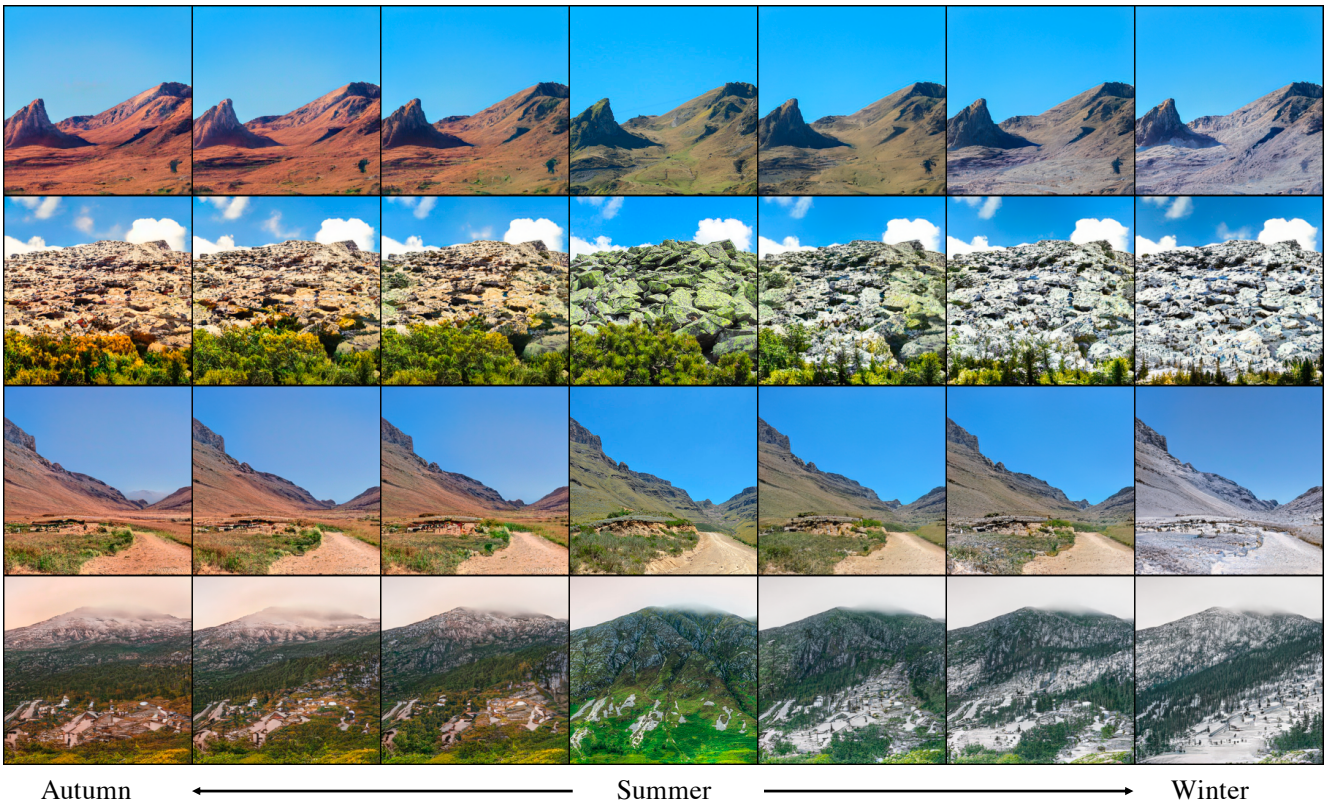


Figure 1: Image attributes editing results of the LSUN mountain dataset [11] under the continuous interpolation. The central column denotes the source summer images, while the remaining columns denote the continuous interpolation images targeting at autumn and winter.

A.2. Unsupervised Semantic Vector Discovery for Image Editing

Another way to conduct image editing is to discover the underlying semantics $\hat{\mathbf{z}}$ via an unsupervised way. Here we adopt the Principal Component Analysis (PCA) [5] to achieve this goal, which could find the orthonormal components in the latent space. Similar to the continuous interpolation approach in our paper, when manipulating the style code using PCA, a good image translation model would keep the content of the images as untouched as possible.

As shown in Fig. 2, FDIT is once again demonstrated to be an identity-preserving model. Specifically, the identities are well maintained, while the only facial attributes such as illumination and hair color are changed.

Figure 2: PCA-based image attributes editing results under the CelebA-HQ dataset. The central column denotes the source images, while within the remaining columns denote the interpolation results of the orthonormal components along two directions.

We additionally show results of image editing in the *full latent space* in Figure 3, which displays more variation.

B. Frequency Domain Image Translation Results

We show the image generation results of the autoencoder based FDIT framework on LSUN Church [11], CelebA-HQ [6], Flickr Waterfalls, and LSUN Bedroom [11] in Figure 4. FDIT framework achieves better performance in preserving the shape, which can be observed in the outline of the churches, the layout of the bedrooms, and the scene of the waterfalls.

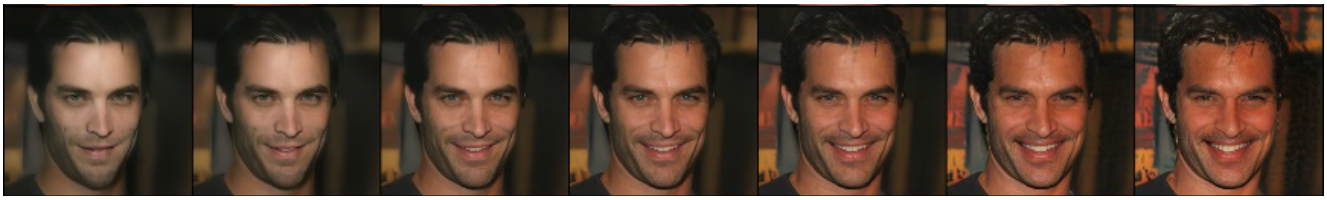
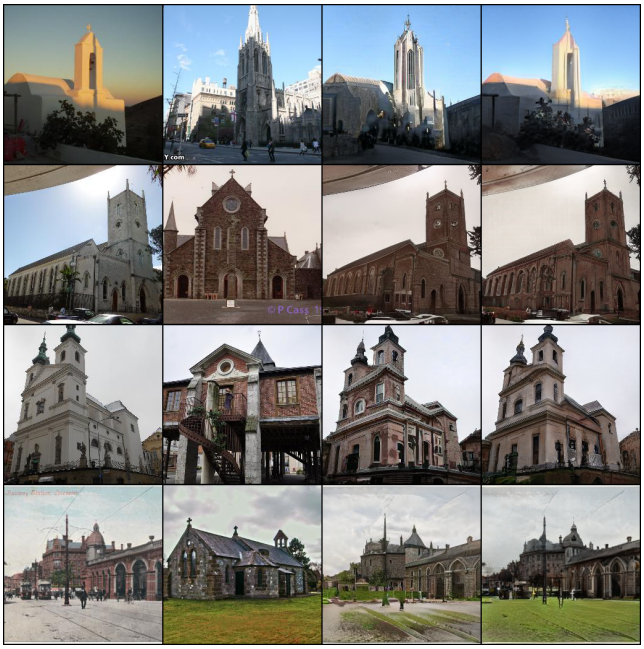


Figure 3: Image editing results using PCA on the full latent space.



(a) LSUN Church

(b) CelebA-HQ

(c) Flicker Waterfalls

(d) LSUN Bedroom

Figure 4: Image translation results under the (a) LSUN Church, (b) CelebA-HQ, (c) Flicker Waterfalls, and (d) LSUN Bedroom dataset. Four columns denote the source images, reference images, and the generated images of Swapping Autoencoder [9] and FDIT, respectively.

Figure 5: Image translation results of the Flicker mountains dataset. From left column to right: we show the source images, reference images, the generated images using Swap AE, with pixel space loss, with Fourier space loss, and with both (FDIT), respectively.

C. Constructing the Flicker Dataset

We collect the large-scale Flicker Mountains dataset and Flicker Waterfalls dataset from `flickr.com`. Each dataset contains 100,000 training images.

D. Training Details

Our Frequency Domain Image Translation (FDIT) framework is composed of the *pixel space* and *Fourier frequency space* losses, which can be conveniently implemented for existing image translation models. For fair comparison, we keep all training and evaluation settings the same as the baselines (Swapping Autoencoder[1] [9], StarGAN v2[2] [2], and Image2StyleGAN[3] [1]). All experiments are conducted on the Tesla V100 GPU.

Swapping Autoencoder [9]. The encoder-decoder backbone is built on StyleGAN2 [7]. We train the model on the 32GB Tesla V100 GPU, where the batch size is 16 for images of 256×256 resolution, and 4 for images of 1024×1024 resolution. During training, a batch of n images are fed into the model, where $\frac{n}{2}$ reconstructed images and $\frac{n}{2}$ image hybrids would be produced. We adopt Adam [8] optimizer where $\beta_1 = 0, \beta_2 = 0.99$. The learning rate is set to be 0.002. The reconstructed quality is supervised by L_1 loss. The discriminator is optimized using the adversarial loss [3]. A patch discriminator is utilized to enhance the texture transferring ability *w.r.t.* reference images.

StarGAN v2 [2]. We use the official implementation in StarGAN v2, where the backbone is built with ResBlocks [4]. The batch size is set to be 8. Adam [8] optimizer is adopted where $\beta_1 = 0, \beta_2 = 0.99$. The learning rate for the encoder, generator, and discriminator is set to be 10^{-4}. In the evaluation stage, we utilize the exponential moving averages over encoder and generator.

Image2StyleGAN v2 [1]. We adopt the Adam optimizer with the learning rate of 0.01, $\beta_1 = 0.9, \beta_2 = 0.999$, and $\epsilon = 1e^{-8}$ in the experiments. We use 5000 gradient descent steps to obtain the GAN-inversion images.

E. Details of Image2StyleGAN and StyleGAN2 results in Table 1.

Both Im2StyleGAN [1] and StyleGAN2 [1] invert the image from the training domain, then use the mixed latent representations to create image hybrids. Image2StyleGAN adopts the iterative optimization on the 'W^+-space' to project images using the StyleGAN-v1 backbone; while StyleGAN2 utilizes an LPIPS-based projector under the StyleGAN-v2 backbone.

F. The qualitative results for Section 4.2

The qualitative results are shown in Figure 5, where FDIT shows better identity preservation than using only pixel or Fourier loss. For example, using only Fourier loss preserves the identity but loses some style consistency in the pixel space.

[1] https://github.com/rosinality/swapping-autoencoder-pytorch
[2] https://github.com/clovaai/stargan-v2
[3] https://github.com/pacifinapacific/StyleGAN_LatentEditor

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. 4

[2] 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. 4

[3] 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. 4

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016. 4

[5] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. In *Proc. NeurIPS*, 2020. 2

[6] 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. 2

[7] 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. 4

[8] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. 4

[9] 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. 3, 4

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

[11] 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. 1, 2