FX-GAN: Self-Supervised GAN Learning via Feature Exchange Supplementary Material

Rui Huang[†]
Carnegie Mellon University
ruih2@alumni.cmu.edu

Wenju Xu[†]
University of Kansas
xuwenju@ku.edu

Teng-Yok Lee Anoop Cherian Ye Wang Tim K. Marks
Mitsubishi Electric Research Laboratories (MERL)
{tlee,cherian,yewang,tmarks}@merl.com

1. Error analysis

Figure 1 shows the errors of the *Exchanged?* task prediction (see Fig. 1 of the paper) when training on ImageNet from iterations 300,000 to 900,000. Note that with self-attention (SAGAN + FX), the error for task reduced to zero more quickly than without self-attention (FX-GAN). This is expected, because the goal of self-attention is to learn to attend to regions that are semantically closely related. Because of this, the inconsistency caused by feature exchange is easier for the discriminator to distinguish, so the proposed feature-exchange loss, ℓ_{fx}, will not be as effective at regularizing the discriminator's representation. For FX-GAN, the error decreases much more slowly, which makes the regularization from ℓ_{fx} more effective and leads to larger improvements in the results. In future work, we could adaptively adjust the difficulty for learning the *Exchanged?* task.

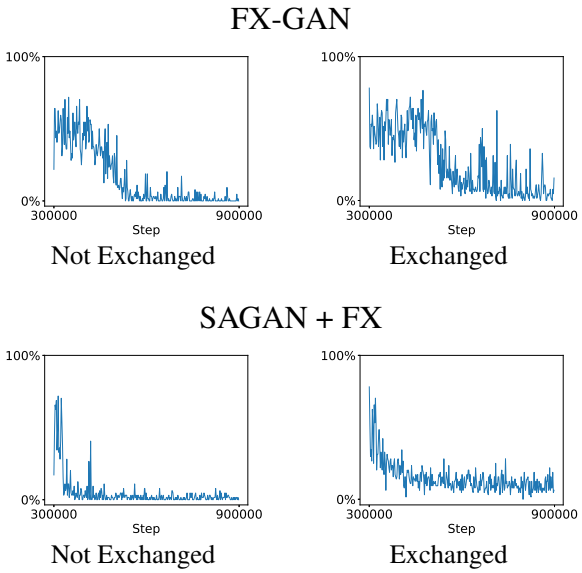


Figure 1. Errors made by the *Exchanged?* task prediction(percentage of images misclassified per batch) when training on ImageNet from iterations 300,000 to 900,000. Errors in the left column are images whose features were *not* exchanged but were misclassified as exchanged. Errors in the right column are images whose features were exchanged but were misclassified as *not* exchanged. *Top row:* FX-GAN. *Bottom row:* SAGAN + FX.

2. Network architecture

For datasets ImageNet, CelebA-HQ, and LSUN bedroom, our network architecture is the same as SAGAN [1]. In the discriminator, each image is first resized to 128×128 pixels, then passed through a sequence of residual blocks. Each residual block downsamples each spatial dimension by 2 and expands the number of channels. Table 1(a) describes the discriminator network architecture by giving the size of the tensor in the spatial and channel dimensions, at the input to the network and after each residual block. For example, the input to the discriminator is a 128×128-pixel image with 3 channels. For the generator, the input noise is first converted into a tensor of $4 \times 4 \times 1024$ elements, then passed through a sequence of deconvolution filters to increase the spatial size and reduce the number of channels. Table 1(b) lists the size of the tensor after each deconvolution. For CIFAR10, since the input size is smaller (32×32), we adjust the network architecture to have fewer residual blocks and fewer deconvolution layers, as described in Table 2.

3. Qualitative results from FX-GAN versus DG-SNGAN

In this supplementary material, we present example images that we generated using the two models that are eval-

[†]Work done while interning at MERL.

Dimension	Input size	Size after each residual block						
x, y	128	64	32	16	8	4	2	2
channels	3	64	128	256	512	1024	2048	2048

(a) Discriminator

Dimension	Input size	Size after each deconvolution					
x, y	4	8	16	32	64	128	128
channels	1024	1024	512	256	128	64	3

(b) Generator

Table 1. The network architecture for LSUN-bedroom, CelebA-HQ, and ImageNet. The numbers represent the tensor shapes after the residual blocks of the discriminator (a) and after the deconvolution blocks of the generator (b).

Dimension	Input size	Size after residual blocks			
x, y	32	32	16	8	4
channels	3	64	128	256	512

(a) Discriminator

Dimension	Input size	Size after deconvolutions			
x, y	4	8	16	32	32
channels	256	256	128	64	3

(b) Generator

Table 2. The network architecture for CIFAR10.

uated in the top section of Table 2 of the paper. The first model is the baseline model, DG-SNGAN. The second is our proposed FX-GAN model (a.k.a. DG-SNGAN + FX). Both models were trained for 1,000,000 iterations on ImageNet (1,000 classes) to perform class-conditional generation of 128×128-pixel images.

3.1. Images generated by our FX-GAN model

Figures 2, 3, and 4 show examples of class-conditional image generation by our proposed model, FX-GAN. Each figure shows 64 generated examples of one class. Each of the 64 images was generated using a different random noise vector.

3.2. Interpolated images generated by FX-GAN

In Figure 5, we show example interpolations of class-conditional images generated by FX-GAN. Each row of images contains a separate interpolation corresponding to a particular class. The ends of each row are images generated from different random noise vectors, while the intermediate images are generated from vectors whose values were interpolated linearly between the two noise vectors.

3.3. Qualitative comparison to DG-SNGAN

We qualitatively compare the class-conditional image generation performance of our FX-GAN model vs. the baseline DG-SNGAN model in Figures 6–12. These examples demonstrate subjective improvements in structural consistency, detail, and/or image diversity for FX-GAN. Interestingly, for some classes, as seen in Figures 10, 11, and 12, the DG-SNGAN baseline seems to exhibit some form of mode collapse (reduction), where greatly reduced image diversity is observed. Across all of the classes, we generally observed that FX-GAN was far more resistant to this type of mode collapse.

References

[1] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena. Self-attention generative adversarial networks. *ICML*, 2019.

Figure 2. FX-GAN generated examples for ImageNet class 15, "robin."

Figure 3. FX-GAN generated examples for ImageNet class 914, "yawl."

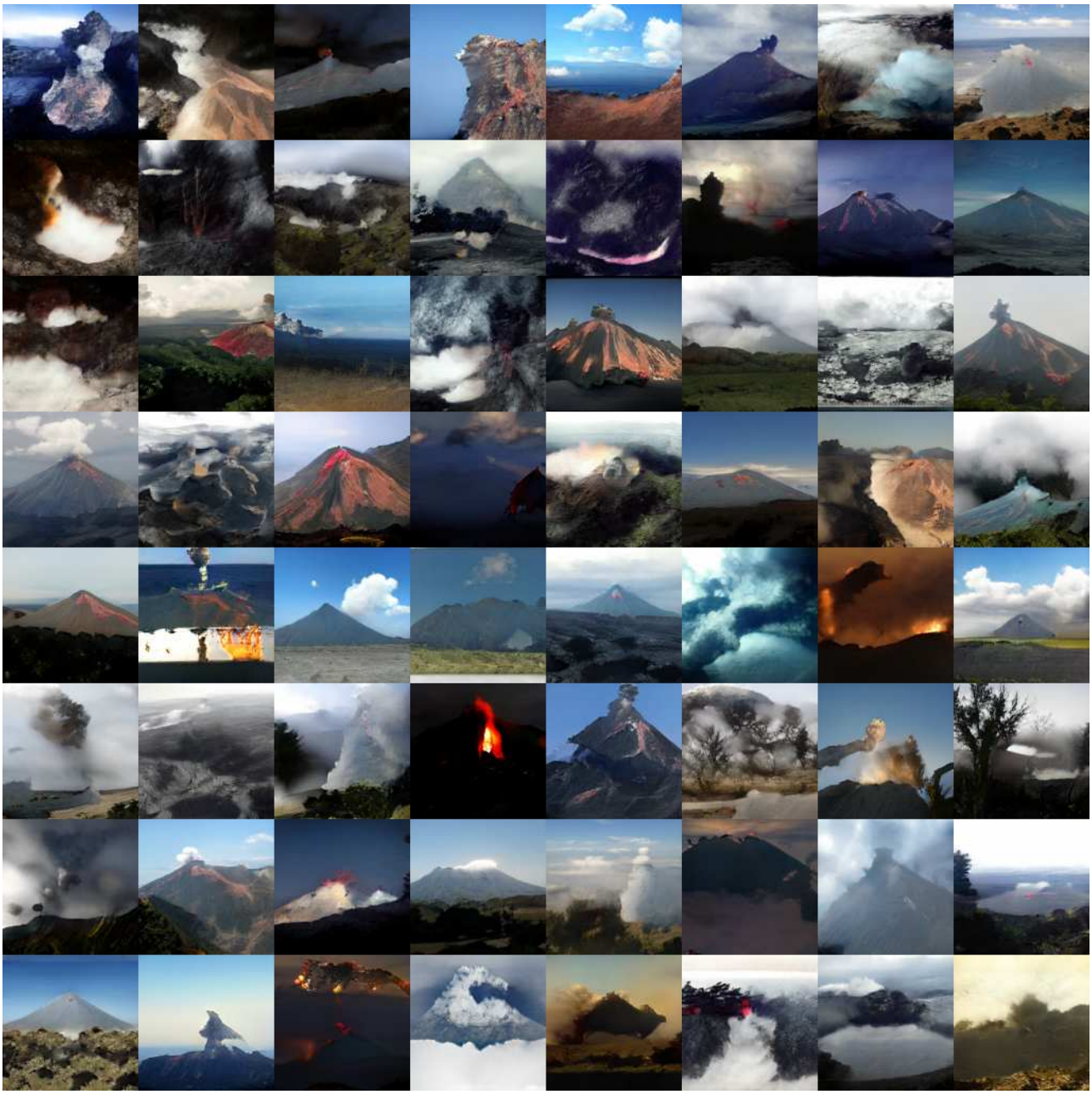


Figure 4. FX-GAN generated examples for ImageNet class 980, "volcano."

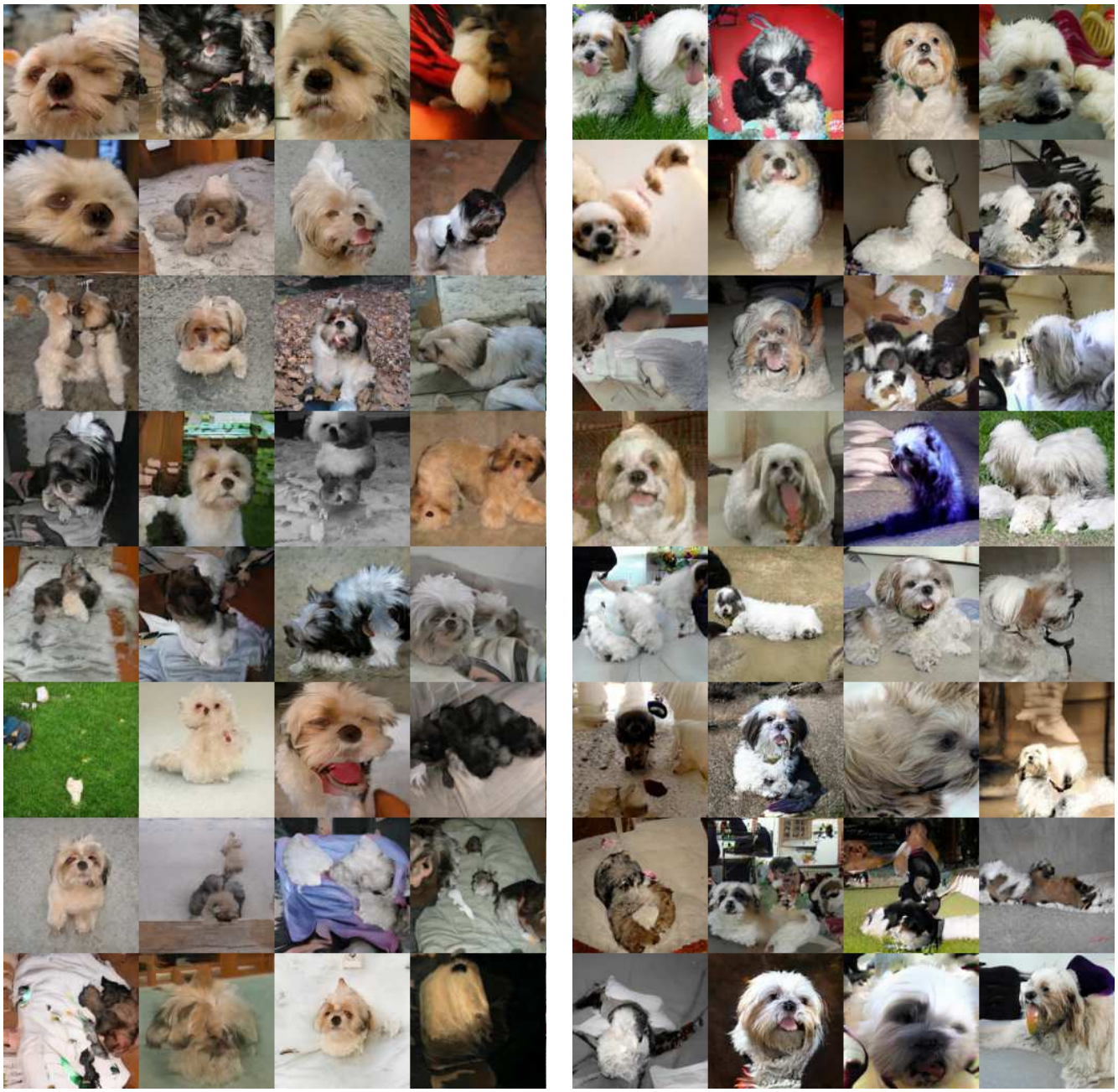
Figure 5. FX-GAN generated image interpolation examples, one class per row. The images on the left and right ends of each row are generated from random noise vectors. The intermediate images in each row are generated from vector values that were linearly interpolated between the two end vectors.

DG-SNGAN FX-GAN

Figure 6. Comparison of DG-SNGAN vs. FX-GAN generated examples for ImageNet class 24, "great grey owl." Note that FX-GAN has learned to generate more realistic eyes than the baseline method.

DG-SNGAN FX-GAN

Figure 7. Comparison of DG-SNGAN vs. FX-GAN generated examples for ImageNet class 81, "ptarmigan." Note that FX-GAN has learned to generate more realistic body shapes than the baseline method.



DG-SNGAN FX-GAN

Figure 8. Comparison of DG-SNGAN vs. FX-GAN generated examples for ImageNet class 155, "Shih-Tzu." Note that FX-GAN has learned to generate more realistic facial arrangements than the baseline method.

DG-SNGAN FX-GAN

Figure 9. Comparison of DG-SNGAN vs. FX-GAN generated examples for ImageNet class 574, "golf ball." Note that FX-GAN has learned to generate more realistic golf ball colors and textures than the baseline method.

DG-SNGAN FX-GAN

Figure 10. Comparison of DG-SNGAN vs. FX-GAN generated examples for ImageNet class 323, "monarch butterfly." Note that FX-GAN has learned to generate better details and color variations than the baseline method.

DG-SNGAN FX-GAN

Figure 11. Comparison of DG-SNGAN vs. FX-GAN generated examples for ImageNet class 520, "crib." Note that FX-GAN creates a much greater variety of crib styles, textures, and colors.

DG-SNGAN FX-GAN

Figure 12. Comparison of DG-SNGAN vs. FX-GAN generated examples for ImageNet class 624, "library." Note that FX-GAN creates a much greater variety of bookshelf styles, textures, and colors.