Appendix: Feature Quantization Improves GAN Training

A. BigGAN

A.1. More results on Ablation Study

In Figure 8, we provide the detailed learning curves under different FQ settings on CIFAR100.

A.2. Experiment setup

- CIFAR-10 and CIFAR-100 (32×32): $bs = 64, ch = 64$. The architecture is given in Table 8. Parameters are set as: $bs = 64, G\_lr = 2e^{-4}, D\_lr = 2e^{-4}, D\_step = 4, G\_step = 1$. To get the best results shown in Table 2, we set $P = 10, \lambda = 0.9, \alpha = 1.0$ of FQ being added at the layers [0, 1, 2, 3].

- ImageNet (64×64): $bs = 512, ch = 64$. The architecture is the same as that in Imagenet (128×128) when you omit the bottom downsample ResBlock in the discriminator and the top upsample ResBlock in the generator, as shown in Table. 10. Parameters are set as: $bs = 512, G\_lr = e^{-4}, D\_lr = 4e^{-4}, D\_step = 1, G\_step = 1$ with self-attention at resolution 32×32. $P = 10, \lambda = 0.7, \alpha = 1.0$ of FQ.

- Imagenet (128×128): The architecture is given in Table. 10. Due to limited hardware resources, compared with the full-version BigGAN, we did the following modification: $bs = 2048 \rightarrow bs = 1024, ch = 96 \rightarrow ch = 64$. $P = 10, \lambda = 0.8, \alpha = 10.0$ of FQ.s

A.3. Generated image samples

We show the generated images for CIFAR-100 in Figure 9, and ImageNet in Figure 10. More high-fidelity results are shown in Figure 11 and Figure 12.

B. StyleGAN

The official discriminator architectures used in StyleGAN and StylgeGAN2 are shwon in Table 9. To apply the FQ techniqure, we did the following minimal modifications:

FQ-StyleGAN In experiments on resolution $32^2 - 128^2$, we put the FQ layer just after Blocks-8 and $P = 10, \lambda = 0.8, \alpha = 1.0$ of FQ. In experiments on resolution 1024^2, the FQ layers were put in Blocks-(16, 32) and $P = 7, \lambda = 0.9, \alpha = 0.25$. Randomly selected samples are shown in Figure 13.

FQ-StylgeGAN2 We put the FQ layer in Blocks-(16, 32) and $P = 7, \lambda = 0.8, \alpha = 0.25$ of FQ. Randomly selected samples are shown in Figure 14.

C. U-GAT-IT

C.1. Dataset

selfie2anime It is first introduced in (Kim et al., 2020). The selfie and anime datasets each contains 3400 training images and 100 testing images.

horse2zebra and photo2vangogh These datasets are used in (Zhu et al., 2017). The training dataset size of each class: 1,067 (horse), 1,334 (zebra), 6,287 (photo), and 400 (vangogh). The test datasets consist of 120 (horse), 140 (zebra), 751 (photo), and 400 (vangogh).

cat2dog and photo2portrait These datasets are used in DRIT (Lee et al., 2018). The numbers of data for each class are 871 (cat), 1,364 (dog), 6,452 (photo), and 1,811 (vangogh). Follow (Kim et al., 2020), we use 120 (horse), 140 (zebra), 751 (photo), and 400 (vangogh) randomly selected images as test data, respectively.

C.2. Architecture

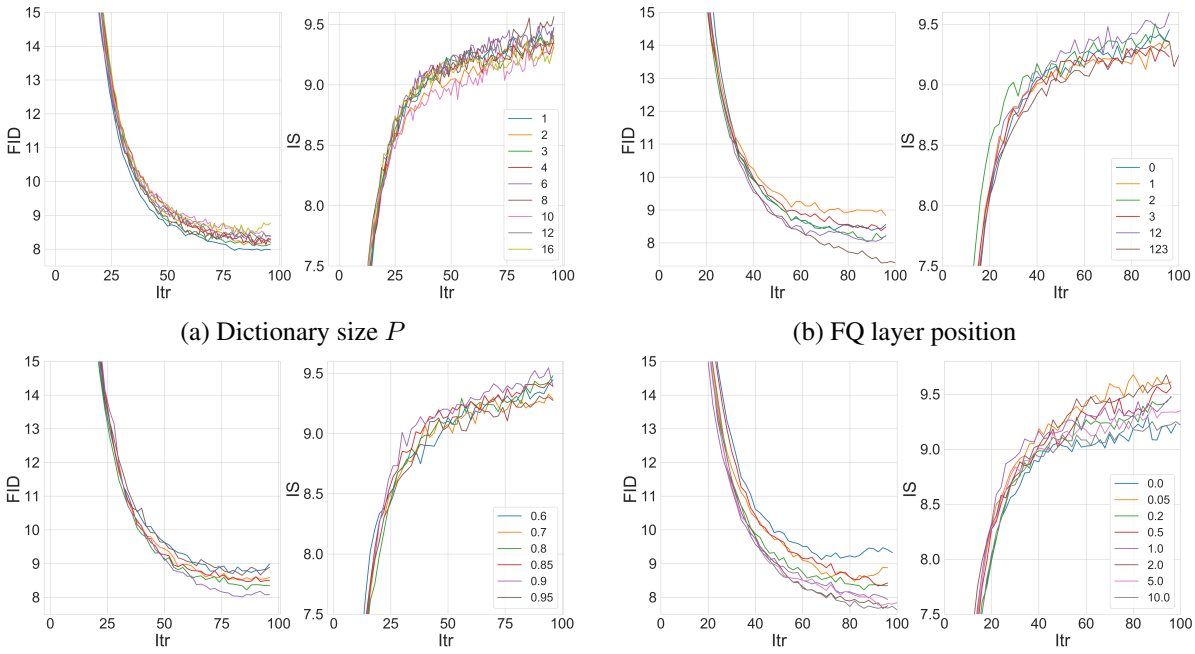
In brief, the U-GAT-IT consists of a generator, a global discriminator and a local discriminator for source to target domain translation and vice versa. We only inject our FQ into the global discriminator and keep other parts unchanged. Training settings are the same as U-GAT-IT. The modified global discriminator architecture is shown in Table 11 and $P = 8, \lambda = 0.8, \alpha = 1.0$ of FQ.

C.3. Additional results

We show more translated images: selfie2anime and anime2selfie in Figure 15, cat2dog and dog2cat in Figure 16, photo2portrait and portrait2photo in Figure 17, vangogh2photo and photo2vangogh in Figure 18, horse2zebra and zebra2horse in Figure 19.

C.4. AMT interface design

The webpage interface used for human evaluation is shown in Figure 20.



(a) Dictionary size P (b) FQ layer position

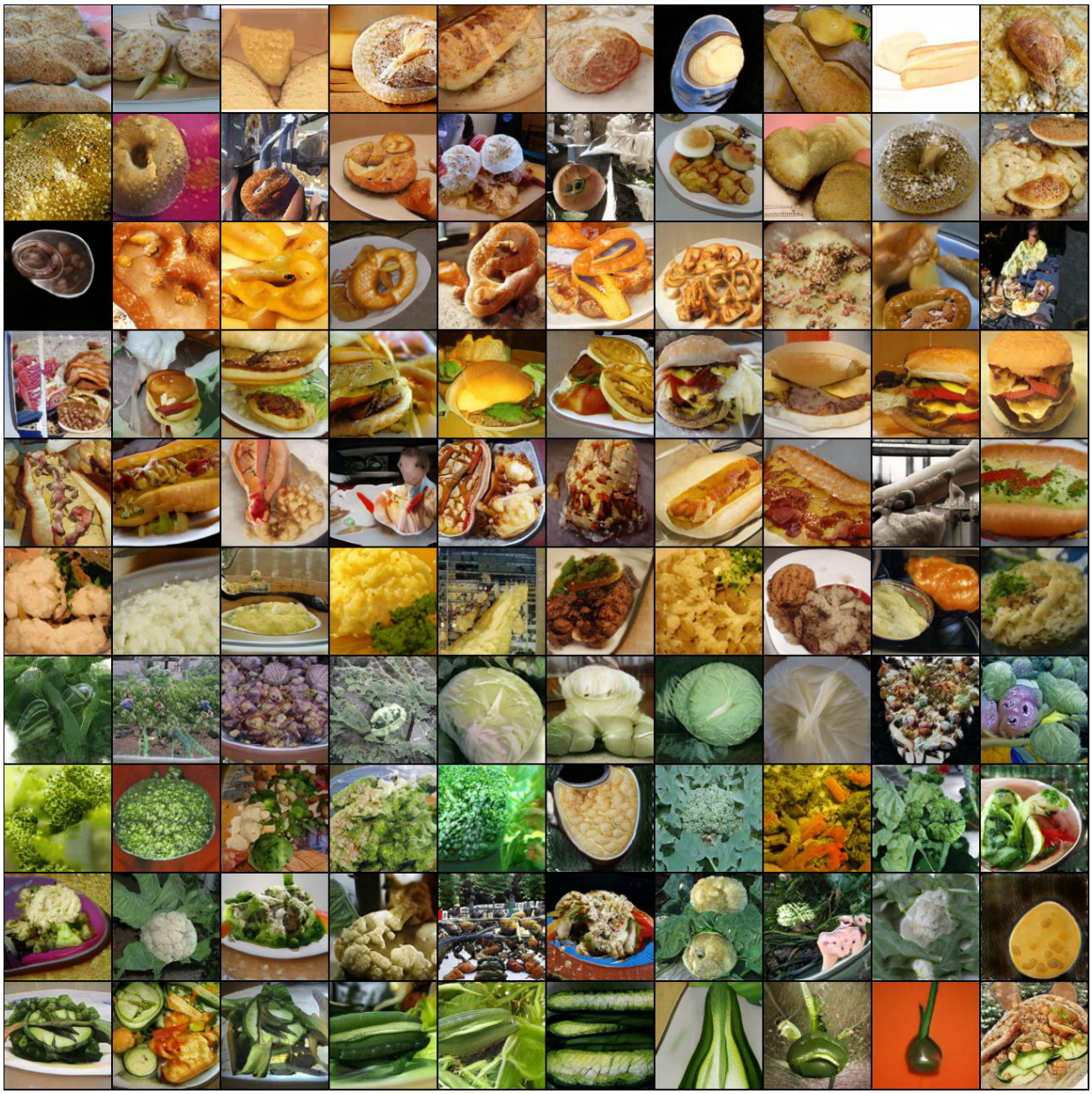
(c) λ (d) α

Figure 8. Ablation studies on the impact of hyper-parameters. The image generation quality is measured with FID ↓ and IS ↑. (a) Dictionary size $K = 2^P$. (b) The positions to apply FQ. (c) The decay hyper-parameter λ in momentum-based dictionary update. (d) The weight α to incorporate FQ,.

Figure 9. Conditionally generated samples (under lowest FID) of BigGAN and FQ-BigGAN on CIFAR-100. (**Top** BigGAN, **Bottom** FQ-BigGAN). FQ-BigGAN obviously surpasses the BigGAN in sample diversity and fidelity.

(a) BigGAN (b) FQ-BigGAN

Figure 10. Conditionally generated samples of BigGAN and FQ-BigGAN on ImageNet. FQ-BigGAN can generate more diverse and accurate samples than BigGAN.



v

Figure 11. More conditionally generated samples of FQ-BigGAN on ImageNet.

Figure 12. More conditionally generated samples of FQ-BigGAN on ImageNet.

Figure 13. Images generated with **FQ**-StyleGAN on FFHQ-1024^2.

Figure 14. Images generated with **FQ**-StyleGAN2 on FFHQ-1024^2.

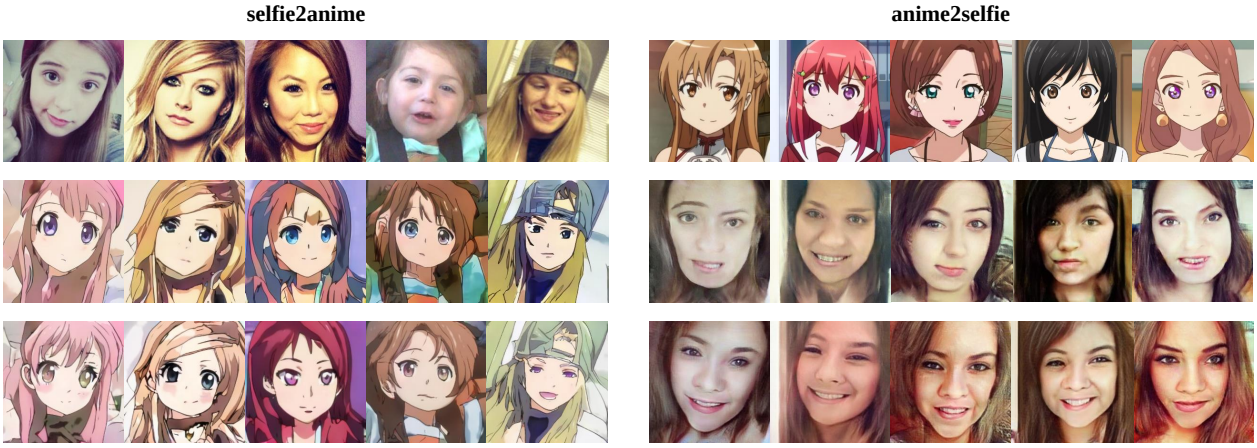


Figure 15. Visual comparisons on selfie2anime and anime2selfie. **First row**: input images. **Second row**: images generate by U-GAT-IT. **Third row**: images generated by FQ-U-GAT-IT.

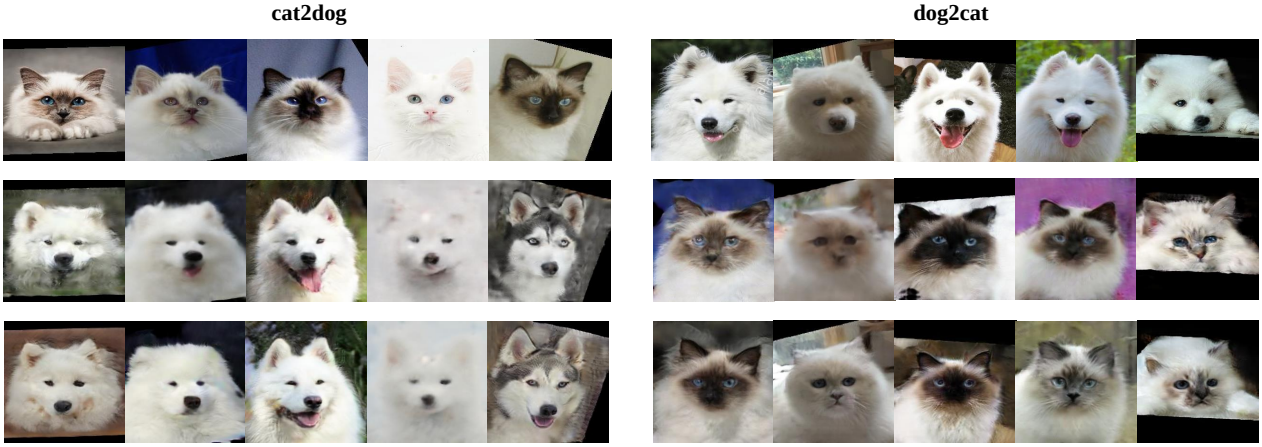


Figure 16. Visual comparisons on cat2dog and dog2cat. **First row**: input images. **Second row**: images generated by U-GAT-IT. **Third row**: images generated by FQ-U-GAT-IT.

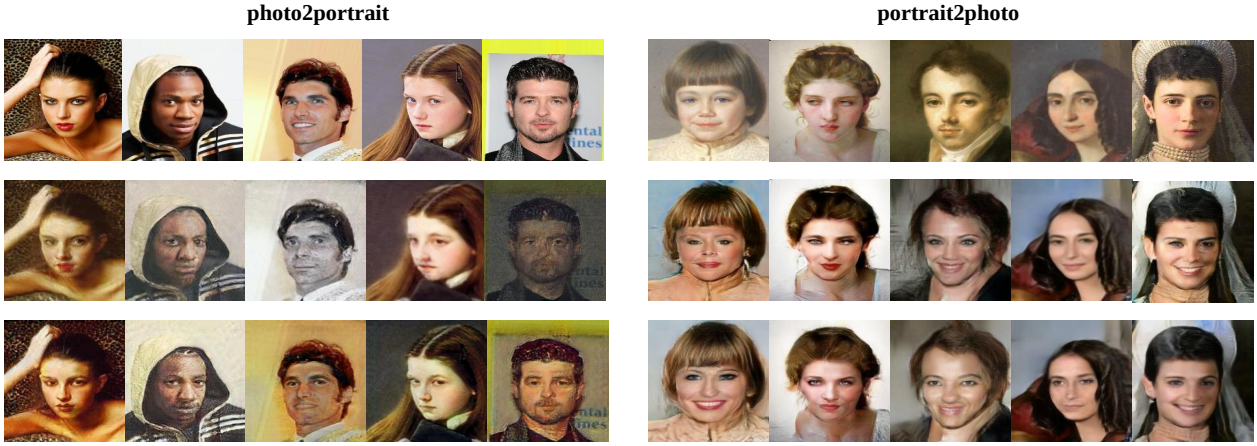


Figure 17. Visual comparisons on photo2portrait and portrait2photo. **First row**: input images. **Second row**: images generated by U-GAT-IT. **Third row**: images generated by FQ-U-GAT-IT.

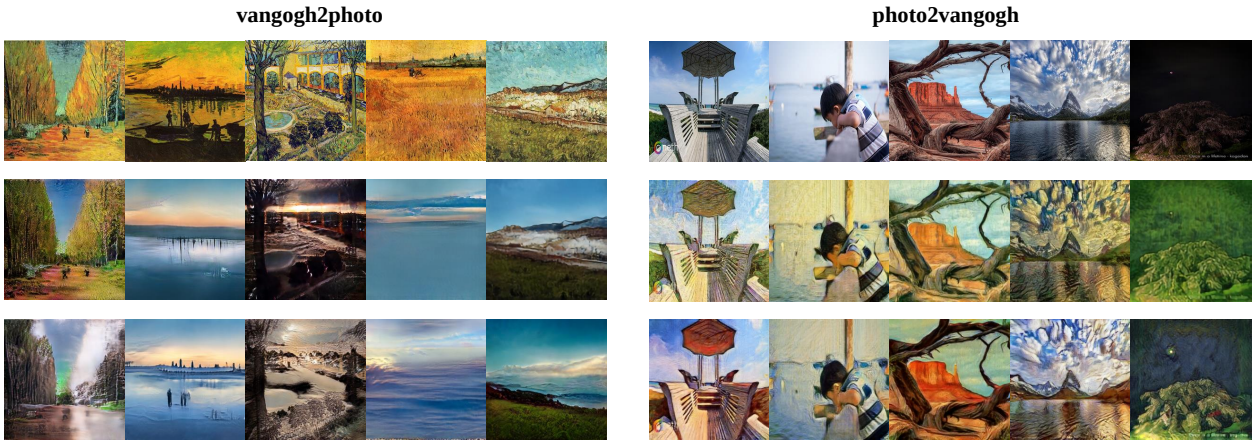


Figure 18. Visual comparisons on vangogh2photo and photo2vangogh. **First row**: input images. **Second row**: images generated by U-GAT-IT. **Third row**: images generated by FQ-U-GAT-IT.

Figure 19. Visual comparisons on horse2zebra and zebra2horse. **First row**: input images. **Second row**: images generated by U-GAT-IT. **Third row**: images generated by FQ-U-GAT-IT. For the horse2zebra translation, U-GAT-IT tends to focus on the texture of zebra but corrupt most details. On contrast, FQ-U-GAT-IT focuses on the horse itself and protect other details. So, FQ-U-GAT-IT fails in some cases (the 4th column) but owns a low KID value.

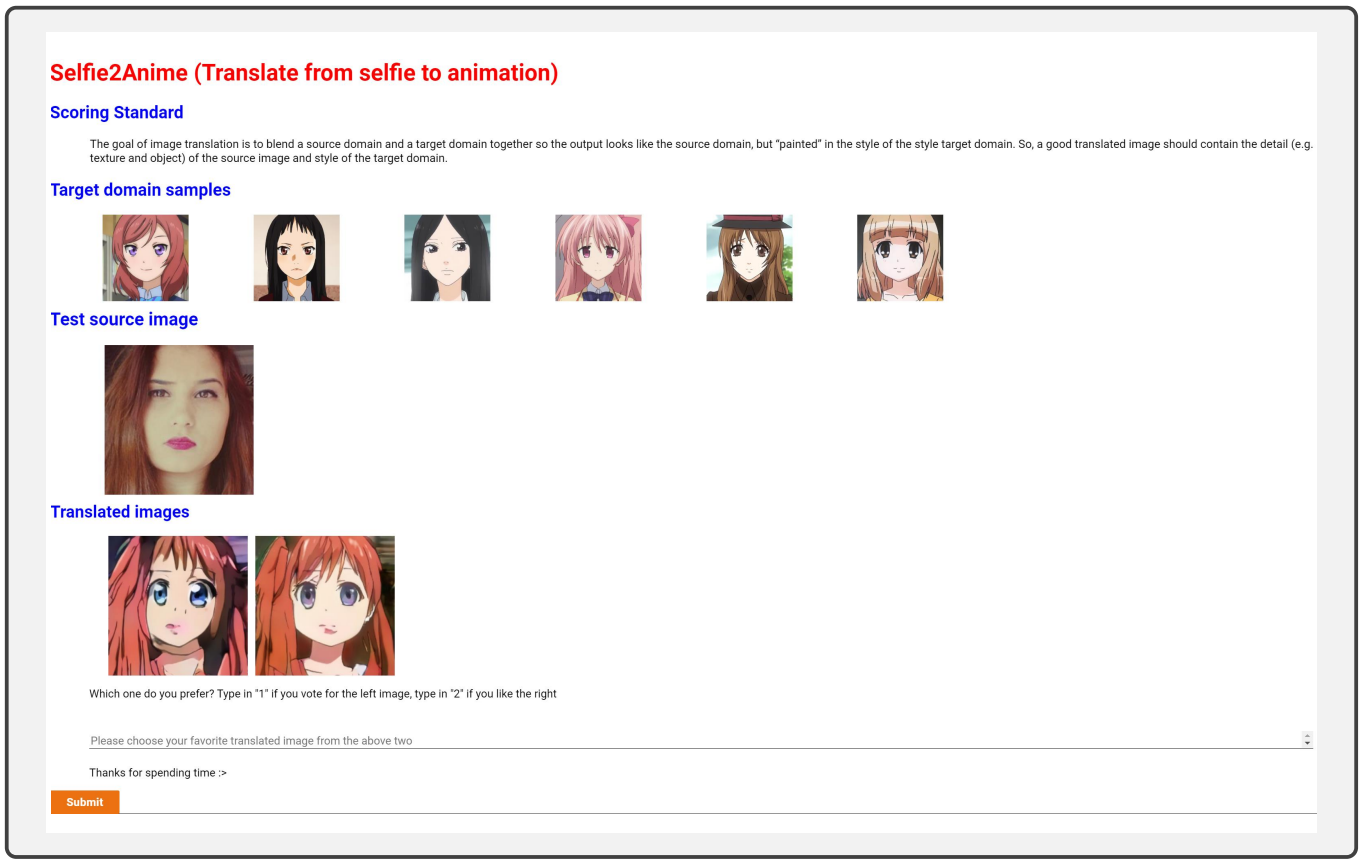


Figure 20. Interface used for human perceptual study on AMT.

Table 8. BigGAN architecture for 32×32 images, $\texttt{ch} = 64$. FQ has been added into different ResBlock layers of discriminator.

$z \in \mathbb{R}^{120} \sim \mathcal{N}(0, 1)$ Embed$(y) \in \mathbb{R}^{128}$
Linear $(20 + 128) \rightarrow 4 \times 4 \times 16ch$
ResBlock up $4ch \rightarrow 4ch$
ResBlock up $4ch \rightarrow 4ch$
ResBlock up $4ch \rightarrow 4ch$
BN, ReLU, 3×3 Conv $ch \rightarrow 3$
Tanh

(a) **Generator**

RGB image $x \in \mathbb{R}^{32 \times 32 \times 3}$
Non-Local Block (64×64)
ResBlock down $4ch \rightarrow 4ch$
ResBlock down $4ch \rightarrow 4ch$
ResBlock $4ch \rightarrow 4ch$
ResBlock $4ch \rightarrow 4ch$
ReLU, Global sum pooling
Embed$(y)\cdot\mathbf{h}$ + (linear $\rightarrow 1$)

(b) **Discriminator**

Table 9. Discriminator architecture in StyleGAN and StyleGAN2

Blocks-#	Input $\rightarrow$ Output shape
1024	$(1024, 1024, 3) \xrightarrow{Conv} (512, 512, 32)$
512	$(512, 512, 32) \xrightarrow{Conv} (256, 256, 64)$
256	$(256, 256, 64) \xrightarrow{Conv} (128, 128, 128)$
128	$(128, 128, 128) \xrightarrow{Conv} (64, 64, 256)$
64	$(64, 64, 256) \xrightarrow{Conv} (32, 32, 512)$
32	$(32, 32, 512) \xrightarrow{Conv} (16, 16, 512)$
16	$(16, 16, 512) \xrightarrow{Conv} (8, 8, 512)$
8	$(8, 8, 512) \xrightarrow{Conv} (4, 4, 512)$
4	$(4, 4, 512) \xrightarrow{Conv} (512)$
Output	$(512) \xrightarrow{Dense} (1)$

Table 10. BigGAN architecture for 128×128 images, $ch = 64$.

$z \in \mathbb{R}^{120} \sim \mathcal{N}(0, 1)$ Embed$(y) \in \mathbb{R}^{128}$
Linear $(20 + 128) \rightarrow 4 \times 4 \times 16ch$
ResBlock up $16ch \rightarrow 16ch$
ResBlock up $16ch \rightarrow 8ch$
ResBlock up $8ch \rightarrow 4ch$
ResBlock up $4ch \rightarrow 2ch$
Non-Local Block (64×64)
ResBlock up $2ch \rightarrow ch$
BN, ReLU, 3×3 Conv $ch \rightarrow 3$
Tanh

(a) **Generator**

RGB image $x \in \mathbb{R}^{128 \times 128 \times 3}$
ResBlock up $ch \rightarrow 2ch$
Non-Local Block (64×64)
ResBlock down $2ch \rightarrow 4ch$
$FQ(K = 2^{10}, 4ch)$
ResBlock down $4ch \rightarrow 8ch$
ResBlock down $8ch \rightarrow 16ch$
ResBlock down $16ch \rightarrow 16ch$
ResBlock $16ch \rightarrow 16ch$
ReLU, Global sum pooling
Embed$(y)\cdot\mathbf{h}$ + (linear $\rightarrow 1$)

(b) **Discriminator**

Table 11. Modified global discriminator of U-GAT-IT (CAM: Class activation maps (Zhou et al., 2016))

Parts	Input $\rightarrow$ Output shape
Encoder Down-sampling	$(h, w, 3) \rightarrow (\frac{h}{2}, \frac{w}{2}, 64)$ $(\frac{h}{2}, \frac{w}{2}, 64) \rightarrow (\frac{h}{4}, \frac{w}{4}, 128)$ $(\frac{h}{4}, \frac{w}{4}, 128) \rightarrow (\frac{h}{8}, \frac{w}{8}, 256)$ $FQ(K = 2^{10}, 256)$ $(\frac{h}{8}, \frac{w}{8}, 256) \rightarrow (\frac{h}{16}, \frac{w}{16}, 512)$ $(\frac{h}{16}, \frac{w}{16}, 512) \rightarrow (\frac{h}{32}, \frac{w}{32}, 1024)$ $(\frac{h}{32}, \frac{w}{32}, 1024) \rightarrow (\frac{h}{32}, \frac{w}{32}, 2048)$
CAM of Discriminator	$(\frac{h}{32}, \frac{w}{32}, 1024) \rightarrow (\frac{h}{32}, \frac{w}{32}, 4096)$ $(\frac{h}{32}, \frac{w}{32}, 4096) \rightarrow (\frac{h}{32}, \frac{w}{32}, 2048)$
Classifier	$(\frac{h}{32}, \frac{w}{32}, 2048) \rightarrow (\frac{h}{32}, \frac{w}{32}, 1)$