

---

# Feature Quantization Improves GAN Training

---

Yang Zhao<sup>\*1</sup> Chunyuan Li<sup>\*2</sup> Ping Yu<sup>1</sup> Jianfeng Gao<sup>2</sup> Changyou Chen<sup>1</sup>

## Abstract

The instability in GAN training has been a long-standing problem despite remarkable research efforts. We identify that instability issues stem from difficulties of performing feature matching with mini-batch statistics, due to a fragile balance between the fixed target distribution and the progressively generated distribution. In this work, we propose Feature Quantization (FQ) for the discriminator, to embed both true and fake data samples into a shared discrete space. The quantized values of FQ are constructed as an evolving dictionary, which is consistent with feature statistics of the recent distribution history. Hence, FQ implicitly enables robust feature matching in a compact space. Our method can be easily plugged into existing GAN models, with little computational overhead in training. Extensive experimental results show that the proposed FQ-GAN can improve the FID scores of baseline methods by a large margin on a variety of tasks, including three representative GAN models on 9 benchmarks, achieving new state-of-the-art performance.

## 1. Introduction

Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) are a powerful class of generative models, successfully applied to a variety of tasks such as image generation (Karras et al., 2019a), image-to-image translation (Liu et al., 2017; Zhu et al., 2017; Isola et al., 2017), text-to-image generation (Zhang et al., 2017), super-resolution (Sønderby et al., 2016), domain adaptation (Tzeng et al., 2017) and sampling from unnormalized distributions (Li et al., 2019).

<sup>\*</sup>Equal contribution

<sup>1</sup>Department of Computer Science and Engineering, University at Buffalo, SUNY <sup>2</sup>Microsoft Research, Redmond. Correspondence to: Chunyuan Li <chunyl@microsoft.com>, Changyou Chen <changyou@buffalo.edu>.

Training GANs is a notoriously challenging task, as it involves optimizing a non-convex problem for its Nash equilibrium in a high-dimensional parameter space. In practice, GANs are typically trained via alternatively updating generator and discriminator, using stochastic gradient descent (SGD) based on mini-batches of true/fake data samples. This procedure is often unstable and lacks theoretical guarantees (Salimans et al., 2016). Consequently, training may exhibit instability, divergence or mode collapse (Mescheder et al., 2018). As a result, many techniques to stabilize GAN training have been proposed (Salimans et al., 2016; Miyato et al., 2018; Karras et al., 2019b).

One possible explanation for the instability is that the learning environment for GANs is non-stationary, and previous models rely heavily on the current mini-batch statistics to match the features across different image regions. Since the mini-batch only provides an estimate, the true underlying distribution can only be learned after passing through a large number of mini-batches. This could prevent adversarial learning on large-scale datasets for a variety of reasons: (i) A small mini-batch may not be able to represent true distribution for large datasets, optimization algorithms may have trouble discovering parameter values that carefully search for continuous features to match fake samples with real samples, and these parameterizations may be brittle and prone to failure when applied to previously unseen images. (ii) Increasing the size of the mini-batch can increase the estimation quality, but doing this also loses the computational efficiency obtained by using SGD. (iii) In particular, the distribution of fake samples shifts as the generator changes during training, making the classification task for discriminator evolve over time (Chen et al., 2019; Liang et al., 2018; Zhao et al., 2020; Cong et al., 2020). In such a non-stationary online environment, discriminator can forget previous tasks if it relies on the statistics from the current single mini-batch, rendering training unstable.

In this work, we show that GANs benefit from feature quantization (FQ) in the discriminator. A dictionary is first constructed via moving-averaged summary of features in recent training history for both true and fake data samples. This enables building a large and consistent dictionary on-the-fly that facilitates the online fashion of GAN training. Each dictionary item represents a unique feature prototype of similar image regions. By quantizing continuous features in



Figure 1. The proposed FQ-GAN generates images by leveraging quantized features from a dictionary, rather than producing arbitrary features in a continuous space when judged by the discriminator. The odd columns show images of the same class (real on the top row, fake at the bottom row), whose corresponding quantized feature maps are shown in the right even column, respectively. The dictionary items are visualized in 1D as the color-bar using t-SNE (Maaten & Hinton, 2008). Image regions with similar semantics utilize the same/similar dictionary items. For example, bird neck is in dark red, sky or clear background is in shallow blue, grass is in orange.

traditional GANs into these dictionary items, the proposed FQ-GAN forces true and fake images to construct their feature representations from the limited values, when judged by discriminator. This alleviates the poor estimate issue of mini-batches in traditional GANs.

To better understand what has been learned during the generation process, we visualize the quantized feature maps of the discriminator in FQ-GAN for different images. Some sample images are shown in Figure 1. Image regions with similar semantics utilize the same or similar dictionary items.

The contributions of this paper are summarized as follows: (i) We propose FQ, a simple yet effective technique that can be added universally to yield better GANs. (ii) The effectiveness of FQ is validated with three GAN models on 10 datasets. Compared with traditional GANs, we show empirically that the proposed FQ-GAN helps training converge faster, and often yields performance improvement by a large margin, measured by generated sample quality. The code is released on Github<sup>1</sup>.

## 2. Background

### 2.1. Preliminaries on vanilla GANs

Consider two general marginal distributions  $q(\mathbf{x})$  and  $p(\mathbf{z})$  over  $\mathbf{x} \in \mathcal{X}$  and  $\mathbf{z} \in \mathcal{Z}$ . To generate samples from these random variables, adversarial methods (Goodfellow et al., 2014) provide a sampling mechanism that only requires gradient backpropagation, without the need to specify the conditional densities. Specifically, instead of sampling directly from the desired conditional distribution, the random variable is generated as a deterministic transformation of an independent noise, *e.g.*, a Gaussian distribution. The sampling procedure for conditionals  $\tilde{\mathbf{x}} \sim p_{\theta}(\mathbf{x}|\mathbf{z})$  is carried

out through the following generating process:

$$\tilde{\mathbf{x}} = g_{\theta}(\mathbf{z}), \mathbf{z} \sim p(\mathbf{z}), \quad (1)$$

where  $g_{\theta}(\cdot)$  is the generators, specified as neural networks with parameters  $\theta$ , and  $p(\mathbf{z})$  is specified as a simple parametric distribution, *e.g.*, isotropic Gaussian  $p(\mathbf{z}) = \mathcal{N}(\mathbf{z}; 0, \mathbf{I})$ . Note that (1) implies that  $p_{\theta}(\mathbf{x}|\mathbf{z})$  is parameterized by  $\theta$ , hence the subscripts.

The goal of GAN (Goodfellow et al., 2014) is to match the marginal  $p_{\theta}(\mathbf{x}) = \int p_{\theta}(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z}$  to  $q(\mathbf{x})$ . Note that  $q(\mathbf{x})$  denotes the true distribution of the data, from which we have samples. In order to do the matching, GAN trains a  $\omega$ -parameterized adversarial discriminator network,  $f_{\omega}(\mathbf{x})$ , to distinguish between samples from  $p_{\theta}(\mathbf{x})$  and  $q(\mathbf{x})$ . Formally, the minimax objective of GAN is given by the following expression:

$$\min_{\theta} \max_{\omega} \mathcal{L}_{\text{GAN}} = \mathbb{E}_{\mathbf{x} \sim q(\mathbf{x})} [\log \sigma(f_{\omega}(\mathbf{x}))] + \mathbb{E}_{\tilde{\mathbf{x}} \sim p_{\theta}(\mathbf{x}|\mathbf{z}), \mathbf{z} \sim p(\mathbf{z})} [\log(1 - \sigma(f_{\omega}(\tilde{\mathbf{x}})))], \quad (2)$$

where  $\sigma(\cdot)$  is the sigmoid function.

### 2.2. Pitfall of Continuous Features

Several works have shown that using feature matching as a training objective of GANs can improve model performance. The basic idea is to embed true/fake distributions in a finite-dimensional continuous feature space, and to match them based on their feature statistics using some divergence metrics. One general form of feature matching is based on Integral probability metric (IPM) (Müller, 1997), indexed by the function space  $\mathcal{F}$ , defined as follows:

$$d_{\mathcal{F}}(p, q) = \sup_{f \in \mathcal{F}} |\mathbb{E}_{\tilde{\mathbf{x}} \sim p(\tilde{\mathbf{x}})} f(\tilde{\mathbf{x}}) - \mathbb{E}_{\mathbf{x} \sim q(\mathbf{x})} f(\mathbf{x})| \quad (3)$$

The particular function class  $\mathcal{F}$  determines the probability metric. For example, Mroueh et al. (2017) proposed

<sup>1</sup><https://github.com/YangNaruto/FQ-GAN>

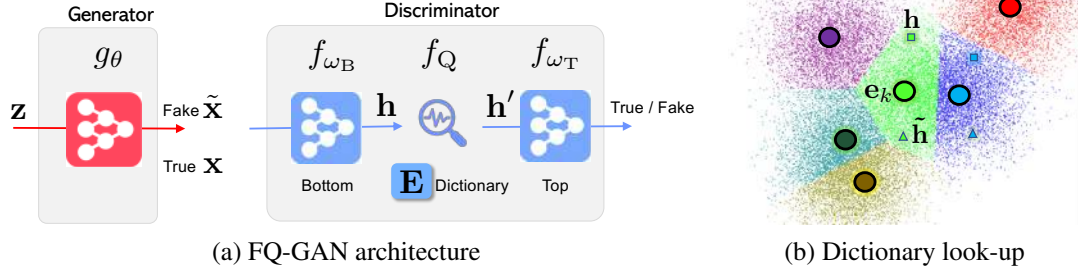


Figure 2. Illustration of FQ-GAN: (a) The neural network architecture. A feature quantization (*i.e.*, dictionary look-up) step  $f_Q$  is injected into the discriminator of the standard GANs. (b) A visualization example of dictionary  $\mathbf{E}$  and the look-up procedure. Each circle “●” indicates a quantization centroid. The true sample features  $\mathbf{h}$  (“■”) and fake sample features  $\tilde{\mathbf{h}}$  (“▲”) are quantized into their nearest centroids  $\mathbf{e}_k$  (represented in the same color in this example), and thus performing implicit feature matching.

MC-GAN, which utilizes both mean and covariance feature statistics. They further showed that several previous works on GAN can be written within the mean feature matching framework, including Wasserstein GAN (Arjovsky et al., 2017), MMD-GAN (Li et al., 2017b), and Improved GAN (Salimans et al., 2016).

Though theoretically attractive, these continuous feature matching methods fall short of recent state-of-the-art performance on large datasets (Brock et al., 2018; Karras et al., 2019a). We argue there are two issues: (i) Principled methods often require to constrain the discriminator capacity (*e.g.*, weight clipping or gradient penalty) to ensure the boundedness: new architectural adjustments such as a higher number of feature maps are needed to compensate for constraints (Mroueh et al., 2017). The question remains what architectural choices can balance the trade-off in practice. (ii) More importantly, the direct feature matching scheme in (3) is estimated via mini-batch statistics, which can be prohibitively inaccurate on large or complex datasets. An effective alternative to match features at large-scale is required, even if it is indirect.

### 3. Feature Quantization GANs

#### 3.1. From Continuous to Quantized Representations

Without loss of generality, the discriminator  $f_\omega(x)$  can be rewritten with a function decomposition:

$$f_\omega(x) = f_{\omega_T} \circ f_{\omega_B}(x), \quad (4)$$

where  $f_{\omega_B}(x)$  is the bottom network, whose output feature  $\mathbf{h} \in \mathbb{R}^D$  is in a  $D$ -dimensional continuous space, and used as the input of the top network  $f_{\omega_T}(\mathbf{h})$ . Instead of working in a continuous feature space, we propose to quantize features into a discrete space, enabling implicit feature matching.

Specifically, we consider the discrete feature space as a dictionary  $\mathbf{E} = \{\mathbf{e}_k \in \mathbb{R}^D \mid k = 1, 2, \dots, K\}$ , where  $K$

is the size of the discrete space (*i.e.*, a  $K$ -way categorical space), and  $D$  is the dimensionality of each dictionary item  $\mathbf{e}_k$ .

The discrete feature  $\mathbf{h}'$  is then calculated by a nearest neighbour look-up using the shared dictionary:

$$\mathbf{h}' = f_Q(\mathbf{h}) = \mathbf{e}_k, \text{ where } k = \underset{j}{\operatorname{argmin}} \|\mathbf{h} - \mathbf{e}_j\|_2, \quad (5)$$

where  $f_Q$  is a parameter-free look-up function, and  $\mathbf{h}'$  is further sent to the top network. Hence, in contrast to the traditional discriminator in (4), our feature quantization discriminator is:

$$f_\omega(x) = f_{\omega_T} \circ f_Q \circ f_{\omega_B}(x), \quad (6)$$

The overall scheme of FQ-GAN is illustrated in Figure 2. FQ-GAN in (6) reduces to the standard GAN model in (4) if  $f_Q$  is removed.

#### 3.2. Dictionary Learning

One remaining question is how to construct the dictionary  $\mathbf{E}$ . Following (Oord et al., 2017), we consider a feature quantization loss consisting of two terms specified in (7): (i) The *dictionary loss*, which only applies to the dictionary items, brings the selected item  $\mathbf{e}$  close to the output of the bottom network. (ii) The *commitment loss* encourages the output of the bottom network to stay close to the chosen dictionary item to prevent it from fluctuating too frequently from one code item to another. The operator  $\operatorname{sg}$  refers to a *stop-gradient* operation that blocks gradients from flowing into its argument, and  $\beta$  is a weighting hyper-parameter ( $\beta = 0.25$  in all our experiments):

$$\mathcal{L}_Q = \underbrace{\|\operatorname{sg}(\mathbf{h}) - \mathbf{e}_k\|_2^2}_{\text{dictionary loss}} + \beta \underbrace{\|\operatorname{sg}(\mathbf{e}_k) - \mathbf{h}\|_2^2}_{\text{commitment loss}}, \quad (7)$$

where  $\mathbf{e}_k$  is the nearest dictionary item to  $\mathbf{h}$  defined in (5).

**Algorithm 1** Feature Quantization GAN

---

**Require:** Randomly initializing the parameters of generator  $g_\theta$ , discriminator  $f_\omega$ , and dictionary  $\mathbf{E}$   
**for** a number of training iterations **do**  
     # Produce a minibatch of true and fake samples  
     Sample  $z \sim p(z)$  and true samples  $x \sim q(x)$ ;  
     Forward  $z$  to generate fake samples  $\tilde{x} = g_\theta(z)$ ;  
     # Feature quantization & Dictionary learning  
     Forward samples  $\{x, \tilde{x}\}$  using (6), and produce  $\mathbf{h}$ ;  
     Feature quantization using (5);  
     Momentum update of dictionary  $\mathbf{E}$  using (8);  
     # Update discriminator  
     Compute gradient  $\frac{\partial \mathcal{L}_{\text{FQGAN}}}{\partial \omega}$  of (9);  
     Update  $\omega$  via gradient ascent;  
     # Update generator  
     Compute gradient  $\frac{\partial \mathcal{L}_{\text{FQGAN}}}{\partial \theta}$  of (9);  
     Update  $\theta$  via gradient descent;  
**end for**

---

**A dynamic & consistent dictionary** The evolution of the generator during GAN training poses a continual learning problem for the discriminator (Liang et al., 2018). In another word, the classification tasks for the discriminator change over time, and recent samples from the generator are more related to current discriminator learning. This inspires us to maintain the dictionary as a queue of features, allowing reusing the encoded features from the preceding mini-batches. The current mini-batch is enqueued to the dictionary, and the oldest mini-batches in the queue are gradually removed. The dictionary always represents a set of prototypes for the recent features, while the extra computation of maintaining this dictionary is manageable. Moreover, removing the features from the oldest mini-batch can be beneficial, because its encoded features are from an early stage of GAN training, and thus the least realistic and consistent with the newest ones.

Alternatively, one may wonder learning a dictionary using all training data beforehand, and keep the dictionary fixed during GAN training. We note this scheme is not practical in that (i) Modern datasets such as ImageNet are usually very large, learning a dictionary offline is prohibitively computationally expensive. (ii) More importantly, such a dictionary is not representative for fake images at the early of training, rendering it difficult to effectively learn quantized features for fake images.

**Momentum update of dictionary.** Specifically, we use the exponential moving average updates to implement the evolving dictionary, as a replacement for the dictionary loss term in (7). For a mini-batch of size  $n$ ,  $n_k$  is the number of features that will be quantized to dictionary item  $e_k$ . The

momentum update for  $e_k$  is:

$$e_k \leftarrow \mathbf{m}_k / N_k, \text{ where } \mathbf{m}_k \leftarrow \lambda \mathbf{m}_k + (1 - \lambda) \sum_{i=1}^{n_k} \mathbf{h}_{i,k},$$

$$N_k \leftarrow \lambda N_k + (1 - \lambda) n_k, \quad (8)$$

where  $\lambda \in (0, 1)$  is a momentum coefficient. Only the parameters in the bottom network  $f_{\omega_B}$  are updated by back-propagation. The momentum updates above make  $e_k$  evolve more smoothly. Small  $\lambda$  considers less history. For example,  $\lambda = 0$  only utilizes the current mini-batch statistics and ignores the entire history, thus (8) reduces to (7). We used the default  $\lambda = 0.90$  in all our experiments.

### 3.3. FQ-GAN Training

The overall training objective of the proposed FQ-GAN is:

$$\min_{\theta, \mathbf{E}} \max_{\omega} \mathcal{L}_{\text{FQ-GAN}} = \mathcal{L}_{\text{GAN}} + \alpha \mathcal{L}_Q, \quad (9)$$

where  $\alpha$  is the weight to incorporate the proposed FQ into GANs. The training procedure is detailed in Algorithm 1. In practice, to avoid degeneration,  $\alpha$  can be annealed from 0 to 1, and  $\mathbf{h}$  (instead of  $\mathbf{h}'$ ) can be used to feed to the next layer at the beginning of training. In this case, one may consider FQ regularizes the learned features using clustering. The generator parameter  $\theta$  and discriminator parameter  $\omega$  are updated via the regularized GAN objective in (9), while the dictionary items  $\mathbf{E}$  are updated via (8). FQ-GAN enjoys several favorable properties, explained as follows.

**Scalability.** The introduction of a dynamic dictionary decouples the dictionary size from the mini-batch size. The dictionary size can be much larger than a typical mini-batch size, and can be flexibly and independently set as a hyper-parameter. The items in the dictionary are progressively replaced. Compared with traditional feature matching methods that only consider the current mini-batch statistics, the proposed FQ-GAN maintains much more representative feature statistics in the dictionary, allowing robust feature matching for large datasets.

**Implicit feature matching.** FQ-GAN shares similar spirits of many other regularization techniques for the discriminator in that they reduce the representational power of the discriminator. However, instead of imposing boundness on weights or gradients, FQ-GAN restricts continuous features into a prescribed set of values, *i.e.*, feature centroids. Since both true and fake samples can only choose their representations from the limited dictionary items, FQ-GAN indirectly performs feature matching. This can be illustrated using the visualization example in Figure 2 (b), where true features  $\mathbf{h}$  and fake features  $\tilde{\mathbf{h}}$  are quantized into the same centroids. Further, the discrete nature improves the possibilities of feature matching, compared to a continuous space.



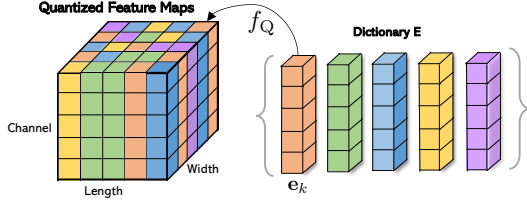


Figure 3. Illustration of FQ construction in CNNs. In this example, the dictionary has 5 items, and feature map is  $\mathbf{h} \in \mathbb{R}^{5 \times 5 \times 5}$ . The feature vector at each position is quantized into a dictionary item, e.g., the back-right feature is quantized into a red item.

### 3.4. FQ-GAN for image generation

FQ is a general method for discriminator design in GANs. We consider image generation tasks in this paper, where the discriminator is often parameterized by convolutional neural networks (CNNs). Each image is represented as a feature map  $\mathbf{h} \in \mathbb{R}^{C \times L \times W}$  in CNNs, where  $C, L, W$  is the number of channels as well as the length and width, respectively. We construct a position-wise dictionary, with each item  $e \in \mathbb{R}^C$ . At a given position on the feature map, the feature vector characterizes the local image region. It is quantized into its nearest dictionary item for calibration, leading to a new quantized feature map  $\mathbf{h}'$  containing calibrated local feature prototypes. We provide the visual illustration on constructing FQ for CNN-based discriminator in Figure 3. Note that the FQ module can be used in multiple different layers of discriminator.

## 4. Related Work

### 4.1. Improving GANs

Training vanilla GANs is difficult: it requires carefully finely-tuned hyper-parameters and network architectures to make it work. Much recent research has accordingly focused on improving its stability, drawing on a growing body of empirical and theoretical insights (Nowozin et al., 2016; Li et al., 2017a; Zhu et al., 2017; Fedus et al., 2017). Among them, the three following aspects are related to FQ-GAN.

**Regularized GANs.** Various Regularization methods have been proposed, including changing the objective functions to encourage convergence (Arjovsky et al., 2017; Mao et al., 2017; Mescheder et al., 2018; Kodali et al., 2017; Zhang et al., 2019b), and constraining discriminator through gradient penalties (Gulrajani et al., 2017) or normalization (Miyato et al., 2018). They counteract the use of unbounded loss functions and ensure that discriminator provides gradients everywhere to generator. FQ is also related to variational discriminator bottleneck (Peng et al., 2018) in the sense that both restrict the feature representation capacity. BigGANs (Brock et al., 2018) use orthogonal regularization, and achieve state of the art image synthesis

performance on ImageNet (Deng et al., 2009).

**Network architectures.** Recent advances consider architecture designs, such as SA-GAN (Zhang et al., 2019a), which adds the self-attention block to capture global structures. Progressive-GAN (Karras et al., 2018) trains high-resolution GANs in the single-class setting by training a single model across a sequence of increasing resolutions. As a new variant, Style-GAN (Karras et al., 2019a) proposed a generator architecture to separate high-level attributes and stochastic variation, achieving highly varied and high-quality human faces.

**Memory-based GANs.** Kim et al. (2018) increased the model complexity via proposing a shared and sophisticated memory module for both generator and discriminator. The generation process is conditioned on samples from the memory. Zhu et al. (2019) proposed a dynamic memory component for image refinement in the text-to-image task.

Compared with the above three aspects, FQ-GAN slightly modifies the discriminator architecture by injecting a dictionary-based look-up layer, and thus regularizes the model capacity to encourage easier feature matching. The dictionary in FQ-GAN can be viewed as a much simpler memory module to store feature statistics. Importantly, our FQ-GAN is easier to use and orthogonal to existing GANs, and can be simply employed as a plug-in module to further improve their performance.

### 4.2. Vector Quantization

Vector quantization (VQ) (Gray, 1984) has been used in various settings, including clustering (Equitz, 1989), metric learning (Schneider et al., 2009), etc. The most related work to ours is (Oord et al., 2017), where discrete latent representations are proposed for variational auto-encoders to circumvent the issues of “posterior collapse”, and show that pairing such quantized representations with an autoregressive prior can generate high-quality images, videos, and speech. Our motivation and scenarios are different from previous VQ works. To the best of our knowledge, this paper presents the first feature quantization work for GANs.

## 5. Experiments

We apply the proposed FQ-GAN method to three state-of-the-art GAN models for a variety of tasks. (i) BigGAN (Brock et al., 2018) for image synthesis, especially for ImageNet, representing a generation task for large-scale datasets. (ii) StyleGAN (Karras et al., 2019a;b) for face synthesis, representing a generation task for high-resolution images. (iii) U-GAT-IT (Kim et al., 2020) for an unsupervised image-to-image translation task.

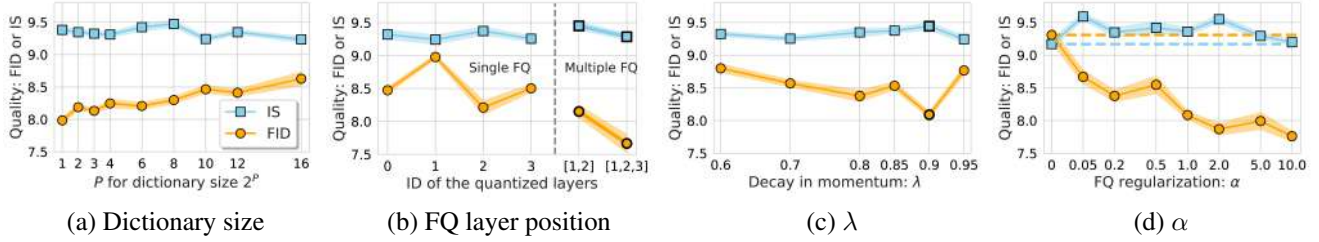


Figure 4. Ablation studies on the impact of hyper-parameters. The image generation quality is measured with FID  $\downarrow$  and IS  $\uparrow$ . (a) Dictionary size  $K = 2^P$ . (b) The positions to apply FQ to discriminator, layer ID is shown on the horizontal axis. (c) The decay hyper-parameter  $\lambda$  in dictionary update. (d) The weight  $\alpha$  to incorporate FQ, the dashed horizon lines are standard GAN baseline  $\alpha = 0$ .

**Evaluation metrics.** We consider three commonly used evaluation metrics for GANs. (i) *Inception Score (IS)* (Salimans et al., 2016) measures how realistic the output of the generator is and the intra-class variety, based on how well the image classification model Inception v3 (Szegedy et al., 2016) classifies them as one of 1,000 known objects collected in ImageNet-1000. Higher scores mean that the model can generate more distinct images. However, it is not reliable when generated images concentrate to the class centers. (ii) *Frchet Inception Distance (FID)* (Heusel et al., 2017) compares the statistics (mean and variances of Gaussian distributions) between the generated samples and real samples. FID is consistent with increasing disturbances and human judgment. Lower scores indicate that the model can generate higher quality images. (iii) *Kernel Inception Distance (KID)* (Bińkowski et al., 2018) improves FID as an unbiased estimator, making it more reliable when there are fewer available test images. We use generated images translated from all test images in the source domain vs. test images in the target domain to compute KID. Lower KID values indicate that images are better translated.

All the baseline methods are implemented via the official codebases from the authors. Model variants that incorporate the proposed feature quantization technique are named with prefix “FQ”. Experiment details are provided in Appendix.

### 5.1. On the impact of hyper-parameters

We investigate the hyper-parameters of FQ on the CIFAR-100 dataset (Krizhevsky et al., 2009). It has 100 classes containing 600 images each, in which there are 500 training images and 100 testing images. Four-layer networks are employed for both the generator and the discriminator. We train the model for 500 epochs, and save a model every 1000 iterations. We take the last 10 checkpoints to report the mean of their performance, measured by FID and IS.

**Dictionary size  $K$ .** In Figure 4 (a), we show the FQ-GAN performance with various dictionary size  $K = 2^P$ . We see that a smaller  $K$  yields better performance. Surprisingly, the dictionary with binary values  $K = 2$  ( $P = 1$ ) provides the best results on this dataset. Larger  $K$  is less favorable for two reasons: (1) it can be more memory expensive; (2) the

Model	FID* $\downarrow$ / IS* $\uparrow$	FID $\downarrow$ / IS $\uparrow$
SN-GAN	14.26 / 8.22	—
R-MMD-GAN	16.21 / 8.29 <sup>†</sup>	—
BigGAN	6.04 / 8.43	6.30 $\pm$ .20 / 8.31 $\pm$ .12
<b>FQ-BigGAN</b>	<b>5.34 / 8.50</b>	<b>5.59<math>\pm</math>.12 / 8.48<math>\pm</math>.03</b>

Table 1. Comparison on CIFAR-10. <sup>†</sup>This number is quoted from (Wang et al., 2019)

method becomes similar to the continuous feature variant, and  $K \rightarrow \infty$  recovers original GANs. Hence, we suggest to choose a smaller  $K$  when the performances are similar.

**Which player/layer to add FQ?** The proposed FQ module can be plugged into either a generator or a discriminator. We found that the performance does not change much when used in the generator. For example, FID is 9.01  $\pm$ .44 and 8.96  $\pm$ .26 before and after adding FQ, respectively. For the discriminator, we place FQ at different positions of the network, and show the results in Figure 4 (b). Multiple FQ layers can generally outperform a single FQ layer.

**Momentum decay  $\lambda$ .** Note that  $\lambda$  determines how much recent history to incorporate when constructing the dictionary. Larger values consider more history. Our experimental results in Figure 4 (c) show that  $\lambda = 0.9$  is a sweet point to balance the current and historical statistics.

**FQ weight  $\alpha$ .** The impact of weighting hyper-parameter  $\alpha$  for FQ in (9) is studied in Figure 4 (d). Adding FQ can immediately improve the baseline by a large margin. Larger  $\alpha$  can further decrease FID while keeping IS values almost unchanged. We used  $\alpha = 1$  for convenience.

### 5.2. BigGAN for Image Generation

BigGAN (Brock et al., 2018) holds the state-of-the-art on the task of class-conditional image synthesis, which benefits from scaling up model size and batch size. Our implementation of BigGAN is based upon BigGAN-PyTorch<sup>2</sup>. We use the same architecture and experimental settings as BigGAN,

<sup>2</sup><https://github.com/ajbrock/BigGAN-PyTorch>

Model	FID* ↓ / IS* ↑	FID ↓ / IS ↑
SN-GAN	16.77 / 7.01	—
TAC-GAN	7.22 / 9.34 <sup>†</sup>	—
<b>FQ-TAC-GAN</b>	<b>7.15 / 9.74</b>	<b>7.21<math>\pm</math>.10 / 9.69<math>\pm</math>.04</b>
BigGAN	8.64 / 9.46	9.01 $\pm$ .44 / 9.36 $\pm$ .10
<b>FQ-BigGAN</b>	<b>7.36 / 9.62</b>	<b>7.42<math>\pm</math>.07 / 9.59<math>\pm</math>.04</b>

Table 2. Comparison on CIFAR-100. <sup>†</sup>This number is quoted from (Gong et al., 2019).

Models	64 × 64	128 × 128
	FID* ↓ / IS* ↑	FID* ↓ / IS* ↑
TAC-GAN	—	23.75 / 28.86 $\pm$ 0.29 <sup>‡</sup>
Half BigGAN	12.75 / 21.84 $\pm$ 0.34	22.77 / 38.05 $\pm$ 0.79 <sup>‡</sup>
<b>FQ-BigGAN</b>	<b>12.62 / 21.99<math>\pm</math>0.32</b>	<b>19.11 / 41.92<math>\pm</math>1.15</b>
256K BigGAN	10.55 / 25.43 $\pm$ 0.15	14.88 / 63.03 $\pm$ 1.42 <sup>†</sup>
<b>FQ-BigGAN</b>	<b>9.67 / 25.96<math>\pm</math>0.24</b>	<b>13.77 / 54.36<math>\pm</math>1.07</b>

Table 3. Comparison on ImageNet-1000 for two resolutions. Both models were trained for 256K iterations if not diverge early. The top and bottom block shows the best results within *half* and *full* of the entire training procedure, respectively. <sup>‡</sup> from (Gong et al., 2019), <sup>†</sup> from (Brock et al., 2018), we cannot reproduce it using their codebase, as the training diverges early.

except for adding FQ layers. Best scores (FID\* / IS\*) and averaged scores (FID / IS) are reported. Standard deviations are computed over five random initializations and their average are reported from the best in each run.

**CIFAR-10** (Krizhevsky et al., 2009) consists of 60K images at resolution  $32 \times 32$  in 10 classes; 50K for training and 10K for testing. 500 epochs are used. The results are show in Table 1. The FQ module improves BigGAN. FQ-BigGAN also outperforms other strong existing GAN models, including spectral normalization (SN) GAN (Miyato et al., 2018), Repulsive MMD-GAN (Wang et al., 2019).

**CIFAR-100** is a more challenging dataset with more fine-grained categories, compared to CIFAR-10. The current best classification model achieves 91.3% accuracy on this dataset (Huang et al., 2019), suggesting that the class distributions have certain support overlaps. 500 epochs are used. We also integrate FQ into TAC-GAN (Gong et al., 2019), which is the current state-of-the-art on CIFAR-100. TAC-GAN improves the intra-class diversity of AC-GAN, thus particularly good at generating images with fine-grained labels. The results are show in Table 2. Experimental settings to achieve these results are provided in A.2. The proposed FQ can improve both TAC-GAN and BigGAN. In particular, FQ significantly improves BigGAN on CIFAR-100 dataset. This is because FQ can increase intra-class diversity, as the dictionary items store a longer distribution history. We show

Model	ImageNet	CIFAR-100	CIFAR-10
BigGAN	7d16h	12h12m	17h37m
<b>FQ-BigGAN</b>	7d19h	12h35m	17h50m

Table 4. Training time comparison of before and after adding FQ module. TITAN XP GPUs are used in these experiments. We train ImageNet ( $64 \times 64$ ) for 256k iterations on 2 GPUs, and CIFAR-100 and CIFAR-10 on 1 GPU for 10k iterations.

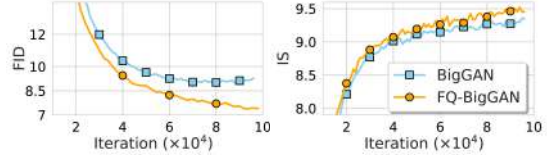


Figure 5. Learning curves on CIFAR-100.

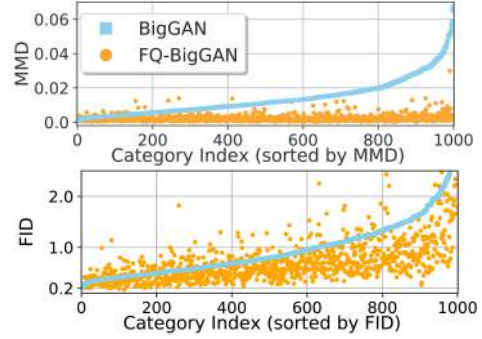


Figure 6. Comparison on per-class metrics for ImageNet.

generated image samples to illustrate the improved diversity for each class in Figure 9 in Appendix.

**ImageNet-1000** (Russakovsky et al., 2015) contains around 1.2 million images with 1000 distinct categories. We pre-process images into resolution  $64 \times 64$  and  $128 \times 128$  in our experiments, respectively. 100 epochs are used. The results are show in Table 3. It shows that FQ improves generation quality for both resolution 64 and 128.

**Computational Cost.** To evaluate the computational overhead of FQ, we compare the running time of BigGAN and our FQ-BigGAN variant on three datasets in Table 4. To finish the same number of training epochs, FQ-GAN takes 1.63%, 3.14%, and 1.23% more time than the original BigGAN on ImageNet, CIFAR-100 and CIFAR-10, respectively. It means that the additional time cost of FQ is negligible. In practice, FQ-GAN converges faster, as shown in Figure 5. It may take less time to reach the same performance.

**How does FQ improve performance?** We perform in-depth analysis on per-class image generation quality on ImageNet. Two metrics are studied: (i) We extract quantized feature sets from discriminator for both real and fake images per class, and measure their distribution divergence



Resolution	$32^2$	$64^2$	$128^2$	$1024^2$
StyleGAN	3.28	4.82	6.33	5.24
<b>FQ-StyleGAN</b>	<b>3.01</b>	<b>4.36</b>	<b>5.98</b>	<b>4.89</b>

Table 5. StyleGAN: Best FID-50k scores in FFHQ at different resolutions.

using maximum mean discrepancy (MMD). Lower MMD values indicate better feature matching. (ii) The per-class FID is also computed from the pre-trained inception network, which measures the matching quality that a generated distribution fits the target distribution in each class. Lower FID values indicate better high intra-class diversity. The results are shown in In Figure 6. FQ yields significantly lower MMD, and lower FID by a large margin, meaning that FQ can improve feature matching, and intra-class diversity.

### 5.3. StyleGAN for Face Synthesis

StyleGAN yields state-of-the-art results in unconditional generative image modeling. StyleGAN (Karras et al., 2019a) is a new variant of the Progressive GAN (Karras et al., 2018), the main difference is the introduction of a latent mapping network in the generator architecture. The very recent version, StyleGAN2 (Karras et al., 2019b), simplifies the progressive architecture, and uses a suite of techniques to improve the performance. We apply our FQ to the StyleGAN and StyleGAN2, based on the TensorFlow codes of StyleGAN<sup>3</sup> and StyleGAN2<sup>4</sup>. The Flickr-Faces-HQ (FFHQ) dataset (Karras et al., 2019a) is used. It consists of 70k high-quality images ( $1024 \times 1024$ ), which endows more variations than the previously widely used CelebA-HQ dataset in terms of accessories, age, ethnicity and image background (Karras et al., 2019a). Each model was trained using 25M images by default. For StyleGAN, we consider four resolutions at  $32^2$ ,  $64^2$ ,  $128^2$  and  $1024^2$ . The progressive training starts from resolution  $8^2$  in experiments of resolution  $32^2 - 128^2$  whereas the initial resolution is  $512^2$  in the experiment on  $1024^2$ . The results are shown in Table 5. The FQ variant improves StyleGAN on all four resolutions. For StyleGAN2, we deploy the model under *config-e* (Karras et al., 2019b) and use the full resolution FFHQ. The best FID score of FQ-StyleGAN2 is **3.19** which surpasses the reported score 3.31 of StyleGAN2. High-fidelity generated faces from the two models are given in Appendix.

### 5.4. Unsupervised Image-to-Image Translation

The task of unsupervised image translation is becoming increasingly popular, inspired by recent advances in GANs. U-GAT-IT (Kim et al., 2020) is the latest state-of-the-art. We validate our FQ using their official TensorFlow code-

base<sup>5</sup>. Five unpaired image datasets are used for evaluation, including selfie2anime (Kim et al., 2020), cat2dog, photo2portrait (Lee et al., 2018), horse2zebra and vangogh2photo (Zhu et al., 2017). All images are resized to  $256 \times 256$  resolution. Details are given in the Appendix.

We also compare with several known image translation models, including CycleGAN (Zhu et al., 2017) and UNIT (Huang et al., 2018), which show better performance than MUNIT (Liu et al., 2017), DRIT (Lee et al., 2018) in (Kim et al., 2020). Each model is trained for 100 epochs, and we report the best KID values in Table 6. FQ improves U-GAT-IT on most datasets, and achieves new state-of-the-art for image translation. We have also conducted human evaluation on Amazon Mechanical Turk (AMT). Each testing image is judged by 3 users, who are asked to select the best translated image to target domain. We inform to the participants the name of target domain, along with six example images of target domain as visual illustration. An example of user interface is show in Figure 20 in Appendix. Table 7 shows the overall percentage that users prefer a particular model. The proposed FQ achieves higher score in human perceptual study (except for comparable results on photo2vangogh), compared to its baseline method. Qualitative comparison in Figure 7 shows that FQ can produce sharper image regions, this is because the dictionary items that FQ utilizes to construct features are from recent history. More examples on translated images are in Appendix.

## 6. Conclusion

In this paper, we propose Feature Quantization Generative Adversarial Networks (FQ-GANs), which incorporate a feature quantization module into the discriminator learning of the GAN framework. The FQ module is effective in performing implicit feature matching for large datasets. FQ can be easily used in training many existing GAN models, and improve their performance. It yields improved performance on three canonical tasks, including BigGAN for image generation on ImageNet, StyleGAN and StyleGAN2 for face generation on FFHQ, and unsupervised image-to-image translation. FQ-GAN sets new state-of-the-art performance on most datasets.

## Acknowledgements

The authors gratefully acknowledge Yanwu Xu for preparing the TAC-GAN codebase, and Yulai Cong for proofreading the draft. We are also grateful to the entire Philly Team inside Microsoft for providing our computing platform.

<sup>5</sup><https://github.com/taki0112/UGATIT>

<sup>3</sup><https://github.com/NVlabs/stylegan>

<sup>4</sup><https://github.com/NVlabs/stylegan2>



Model	selfie2anime	horse2zebra	cat2dog	photo2portrait	photo2vangogh
UNIT	14.71 $\pm$ 0.59	10.44 $\pm$ 0.67	8.15 $\pm$ 0.48	1.20 $\pm$ 0.31	4.26 $\pm$ 0.29
CycleGAN	13.08 $\pm$ 0.49	8.05 $\pm$ 0.72	8.92 $\pm$ 0.69	1.84 $\pm$ 0.34	5.46 $\pm$ 0.33
U-GAT-IT	11.61 $\pm$ 0.57	7.06 $\pm$ 0.8	7.07 $\pm$ 0.65	1.79 $\pm$ 0.34	4.28 $\pm$ 0.33
<b>FQ-U-GAT-IT</b>	<b>11.40 <math>\pm</math> 0.28</b>	<b>2.93 <math>\pm</math> 0.36</b>	<b>6.44 <math>\pm</math> 0.35</b>	<b>1.09 <math>\pm</math> 0.17</b>	6.54 $\pm$ 0.18

Model	anime2selfie	zebra2horse	dog2cat	portrait2photo	vangogh2photo
UNIT	26.32 $\pm$ 0.92	14.93 $\pm$ 0.75	9.81 $\pm$ 0.34	1.42 $\pm$ 0.24	9.72 $\pm$ 0.33
CycleGAN	11.84 $\pm$ 0.74	8.0 $\pm$ 0.66	9.94 $\pm$ 0.36	1.82 $\pm$ 0.36	4.68 $\pm$ 0.36
U-GAT-IT	11.52 $\pm$ 0.57	7.47 $\pm$ 0.71	8.15 $\pm$ 0.66	1.69 $\pm$ 0.53	5.61 $\pm$ 0.32
<b>FQ-U-GAT-IT</b>	<b>10.23 <math>\pm</math> 0.40</b>	<b>7.10 <math>\pm</math> 0.42</b>	8.90 $\pm$ 0.32	<b>0.73 <math>\pm</math> 0.16</b>	5.21 $\pm$ 0.22

Table 6. KID  $\times 100$  for different image translation datasets. All numbers except for our FQ variant are from (Kim et al., 2020).

Model	baseline	<b>FQ</b>
selfie2anime	44.7	<b>55.3</b>
horse2zebra	36.2	<b>63.8</b>
cat2dog	34.0	<b>66.0</b>
photo2portrait	42.5	<b>57.5</b>
photo2vangogh	48.8	<b>51.2</b>

Table 7. User perceptual study on translated image preference (in percentage) between U-GAT-IT and its FQ variant using AMT.

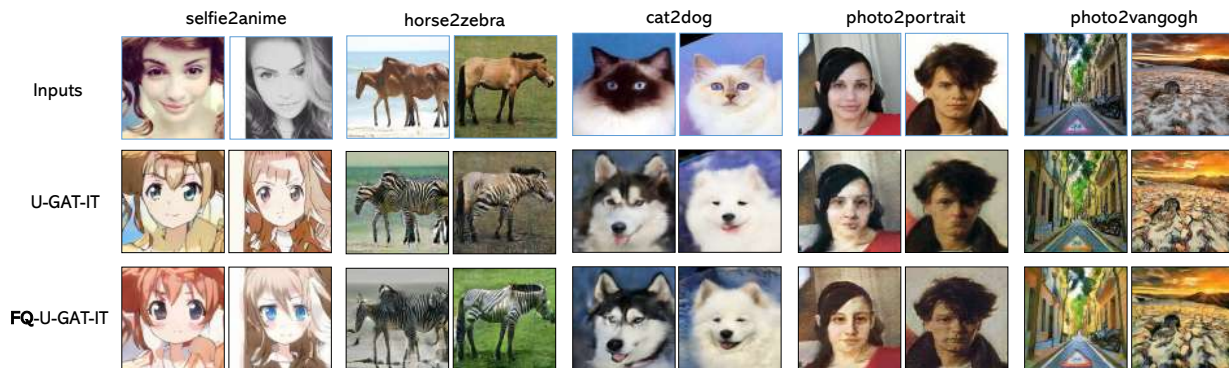


Figure 7. Qualitative comparison. The 1st, 2nd and 3rd shows source and the translated images using U-GAT-IT and FQ, respectively.

## References

- Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein GAN. *arXiv preprint arXiv:1701.07875*, 2017.
- Bińkowski, M., Sutherland, D. J., Arbel, M., and Gretton, A. Demystifying MMD GANs. *arXiv preprint arXiv:1801.01401*, 2018.
- Brock, A., Donahue, J., and Simonyan, K. Large scale GAN training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.
- Chen, T., Lucic, M., Houlsby, N., and Gelly, S. On self modulation for generative adversarial networks. *ICLR*, 2019.
- Cong, Y., Zhao, M., Li, J., Wang, S., and Carin, L. Gan memory with no forgetting. *arXiv preprint arXiv:2002.11810*, 2020.
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.
- Equitz, W. H. A new vector quantization clustering algorithm. *IEEE transactions on acoustics, speech, and signal processing*, 1989.
- Fedus, W., Rosca, M., Lakshminarayanan, B., Dai, A. M., Mohamed, S., and Goodfellow, I. Many paths to equilibrium: Gans do not need to decrease a divergence at every step. *arXiv preprint arXiv:1710.08446*, 2017.
- Gong, M., Xu, Y., Li, C., Zhang, K., and Batmanghelich, K. Twin auxiliary classifiers GAN. In *Advances in Neural Information Processing Systems*, 2019.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.
- Gray, R. Vector quantization. *IEEE Assp Magazine*, 1984.
- Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. Improved training of wasserstein GANs. In *Advances in neural information processing systems*, 2017.
- Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pp. 6626–6637, 2017.

- Huang, X., Liu, M.-Y., Belongie, S., and Kautz, J. Multi-modal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 172–189, 2018.
- Huang, Y., Cheng, Y., Bapna, A., Firat, O., Chen, D., Chen, M., Lee, H., Ngiam, J., Le, Q. V., Wu, Y., et al. Gpipe: Efficient training of giant neural networks using pipeline parallelism. In *Advances in Neural Information Processing Systems*, 2019.
- Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.
- Karras, T., Aila, T., Laine, S., and Lehtinen, J. Progressive growing of GANs for improved quality, stability, and variation. *ICLR*, 2018.
- Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410, 2019a.
- Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. Analyzing and improving the image quality of styleGAN. *arXiv preprint arXiv:1912.04958*, 2019b.
- Kim, J., Kim, M., Kang, H., and Lee, K. U-gat-it: unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. *ICLR*, 2020.
- Kim, Y., Kim, M., and Kim, G. Memorization precedes generation: Learning unsupervised GANs with memory networks. *arXiv preprint arXiv:1803.01500*, 2018.
- Kodali, N., Abernethy, J., Hays, J., and Kira, Z. On convergence and stability of GANs. *arXiv preprint arXiv:1705.07215*, 2017.
- Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.
- Lee, H.-Y., Tseng, H.-Y., Huang, J.-B., Singh, M., and Yang, M.-H. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 35–51, 2018.
- Li, C., Liu, H., Chen, C., Pu, Y., Chen, L., Henao, R., and Carin, L. ALICE: Towards understanding adversarial learning for joint distribution matching. In *Advances in Neural Information Processing Systems*, 2017a.
- Li, C., Bai, K., Li, J., Wang, G., Chen, C., and Carin, L. Adversarial learning of a sampler based on an unnormalized distribution. *AISTATS*, 2019.
- Li, C.-L., Chang, W.-C., Cheng, Y., Yang, Y., and Póczos, B. MMD GAN: Towards deeper understanding of moment matching network. In *Advances in Neural Information Processing Systems*, 2017b.
- Liang, K. J., Li, C., Wang, G., and Carin, L. Generative adversarial network training is a continual learning problem. *arXiv preprint arXiv:1811.11083*, 2018.
- Liu, M.-Y., Breuel, T., and Kautz, J. Unsupervised image-to-image translation networks. In *Advances in neural information processing systems*, pp. 700–708, 2017.
- Maaten, L. v. d. and Hinton, G. Visualizing data using t-SNE. *Journal of machine learning research*, 2008.
- Mao, X., Li, Q., Xie, H., Lau, R. Y., Wang, Z., and Paul Smolley, S. Least squares generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.
- Mescheder, L., Geiger, A., and Nowozin, S. Which training methods for GANs do actually converge? *arXiv preprint arXiv:1801.04406*, 2018.
- Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.
- Mroueh, Y., Sercu, T., and Goel, V. McGAN: Mean and covariance feature matching gan. *arXiv preprint arXiv:1702.08398*, 2017.
- Müller, A. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, pp. 429–443, 1997.
- Nowozin, S., Cseke, B., and Tomioka, R. f-GAN: Training generative neural samplers using variational divergence minimization. In *Advances in neural information processing systems*, 2016.
- Oord, A. v. d., Vinyals, O., and Kavukcuoglu, K. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, 2017.
- Peng, X. B., Kanazawa, A., Toyer, S., Abbeel, P., and Levine, S. Variational discriminator bottleneck: Improving imitation learning, inverse rl, and gans by constraining information flow. *arXiv preprint arXiv:1810.00821*, 2018.
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3): 211–252, 2015.

- Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. Improved techniques for training GANs. In *Advances in neural information processing systems*, pp. 2234–2242, 2016.
- Schneider, P., Biehl, M., and Hammer, B. Distance learning in discriminative vector quantization. *Neural computation*, 2009.
- Sønderby, C. K., Caballero, J., Theis, L., Shi, W., and Huszár, F. Amortised MAP inference for image super-resolution. *arXiv preprint arXiv:1610.04490*, 2016.
- Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.
- Tzeng, E., Hoffman, J., Saenko, K., and Darrell, T. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7167–7176, 2017.
- Wang, W., Sun, Y., and Halgamuge, S. Improving MMD-GAN training with repulsive loss function. *ICLR*, 2019.
- Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., and Metaxas, D. N. StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2017.
- Zhang, H., Goodfellow, I., Metaxas, D., and Odena, A. Self-attention generative adversarial networks. *ICML*, 2019a.
- Zhang, H., Zhang, Z., Odena, A., and Lee, H. Consistency regularization for generative adversarial networks. *arXiv preprint arXiv:1910.12027*, 2019b.
- Zhao, M., Cong, Y., and Carin, L. On leveraging pre-trained gans for limited-data generation. *arXiv preprint arXiv:2002.11810*, 2020.
- Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921–2929, 2016.
- Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2017.
- Zhu, M., Pan, P., Chen, W., and Yang, Y. DM-GAN: Dynamic memory generative adversarial networks for text-to-image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

---

## 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 ( $32 \times 32$ ):  $bs = 64, ch = 64$ . 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 ( $64 \times 64$ ):  $bs = 512, ch = 64$ . The architecture is the same as that in Imagenet ( $128 \times 128$ ) 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  $32 \times 32$ .  $P = 10, \lambda = 0.7, \alpha = 1.0$  of FQ.
- Imagenet ( $128 \times 128$ ): 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 shown in Table 9. To apply the FQ technique, 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.



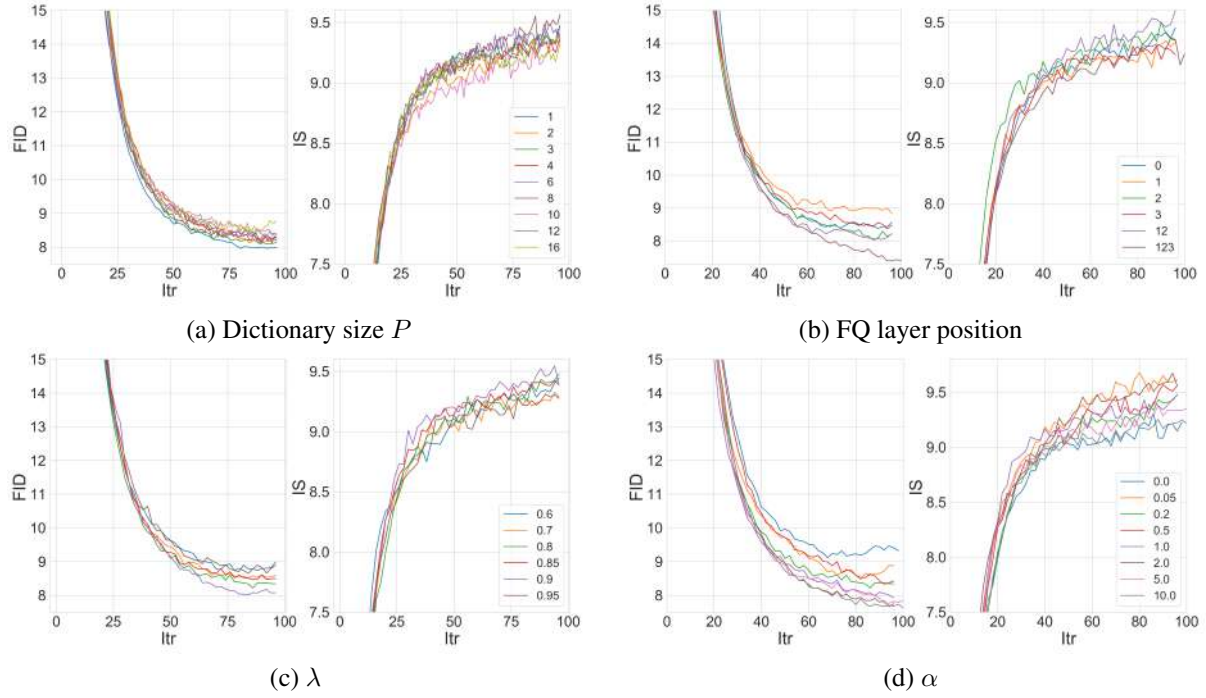


Figure 8. Ablation studies on the impact of hyper-parameters. The image generation quality is measured with FID  $\downarrow$  and IS  $\uparrow$ . (a) Dictionary size  $K = 2^P$ . (b) The positions to apply FQ. (c) The decay hyper-parameter  $\lambda$  in momentum-based dictionary update. (d) The weight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



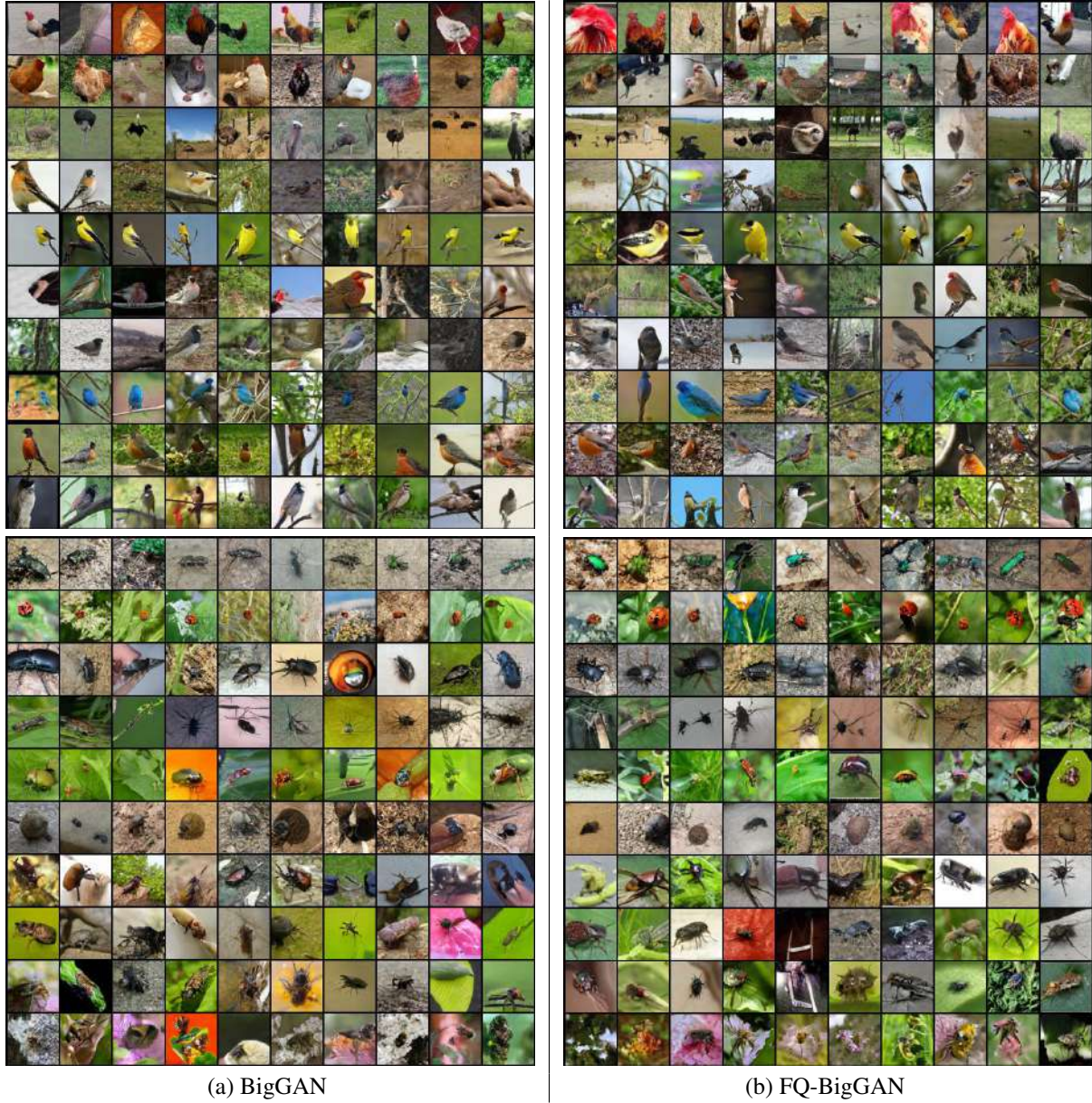
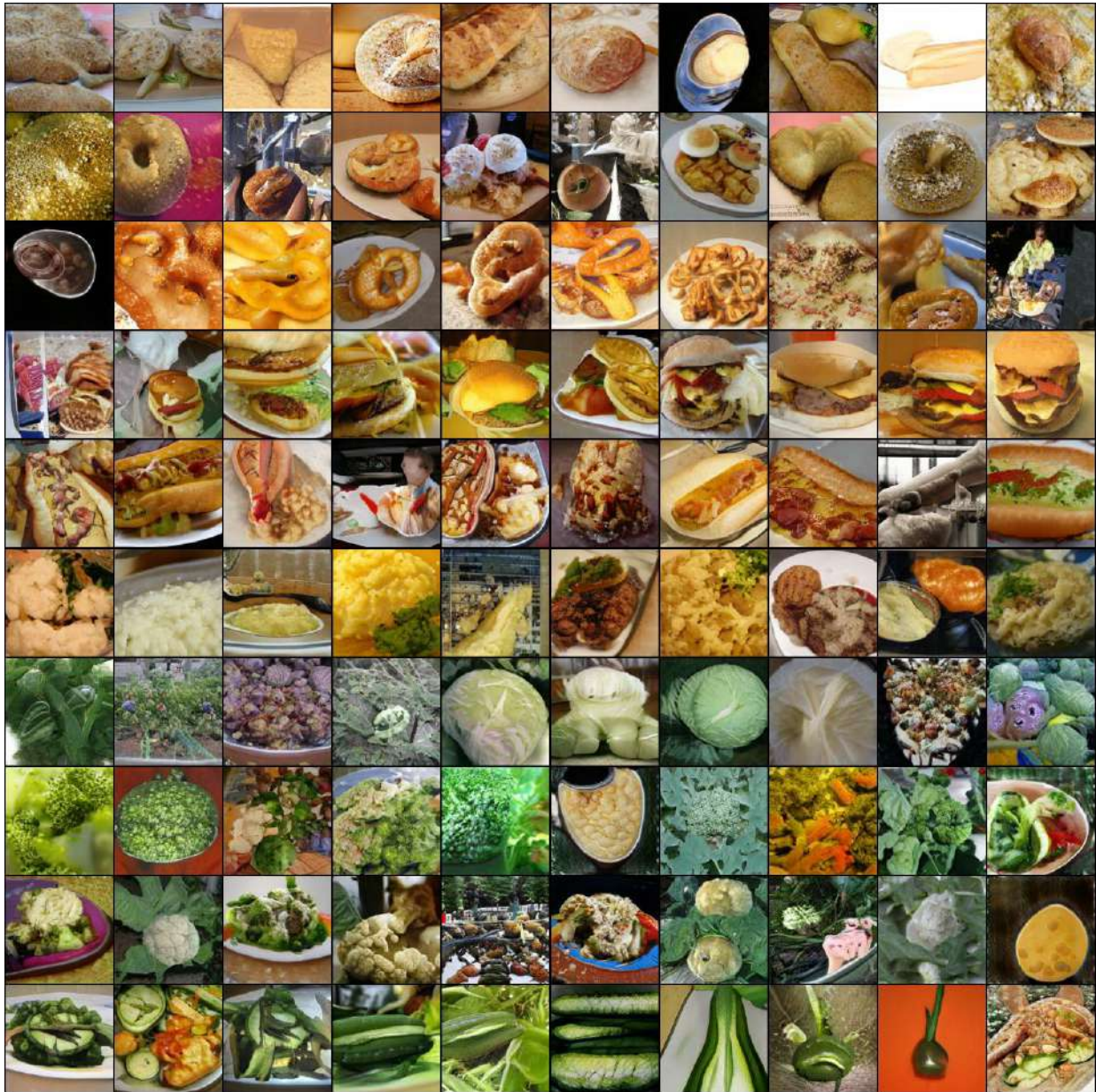


Figure 10. Conditionally generated samples of BigGAN and FQ-BigGAN on ImageNet. FQ-BigGAN can generate more diverse and accurate samples than BigGAN.





v

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



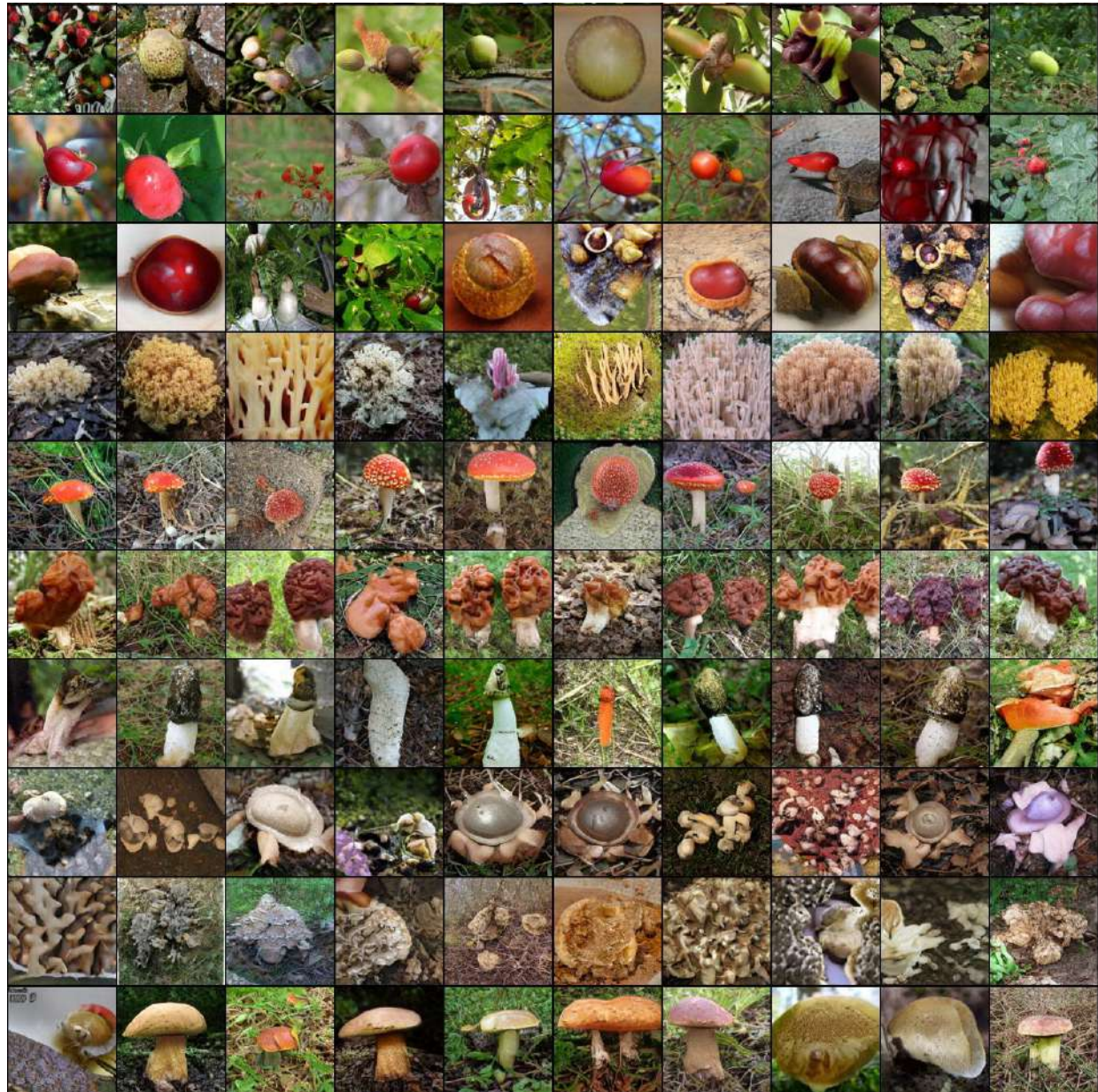


Figure 12. More conditionally generated samples of FQ-BigGAN on ImageNet.



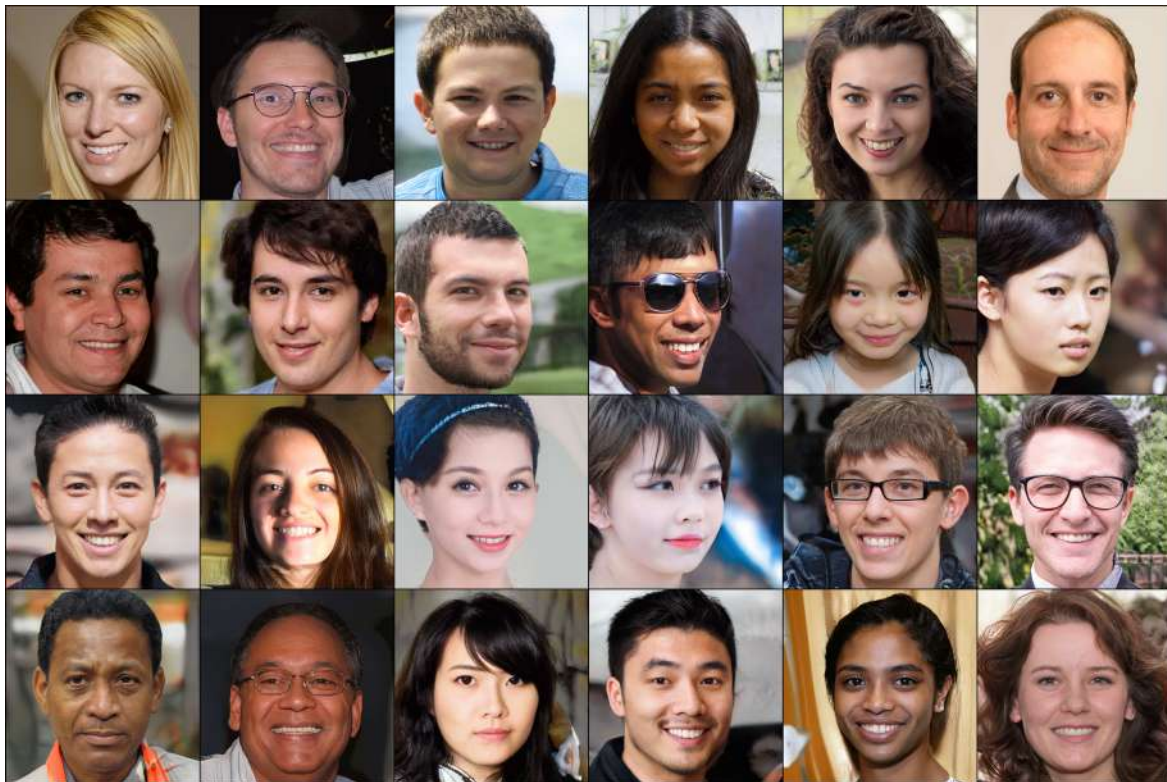


Figure 13. Images generated with **FQ**-StyleGAN on FFHQ-1024<sup>2</sup>.

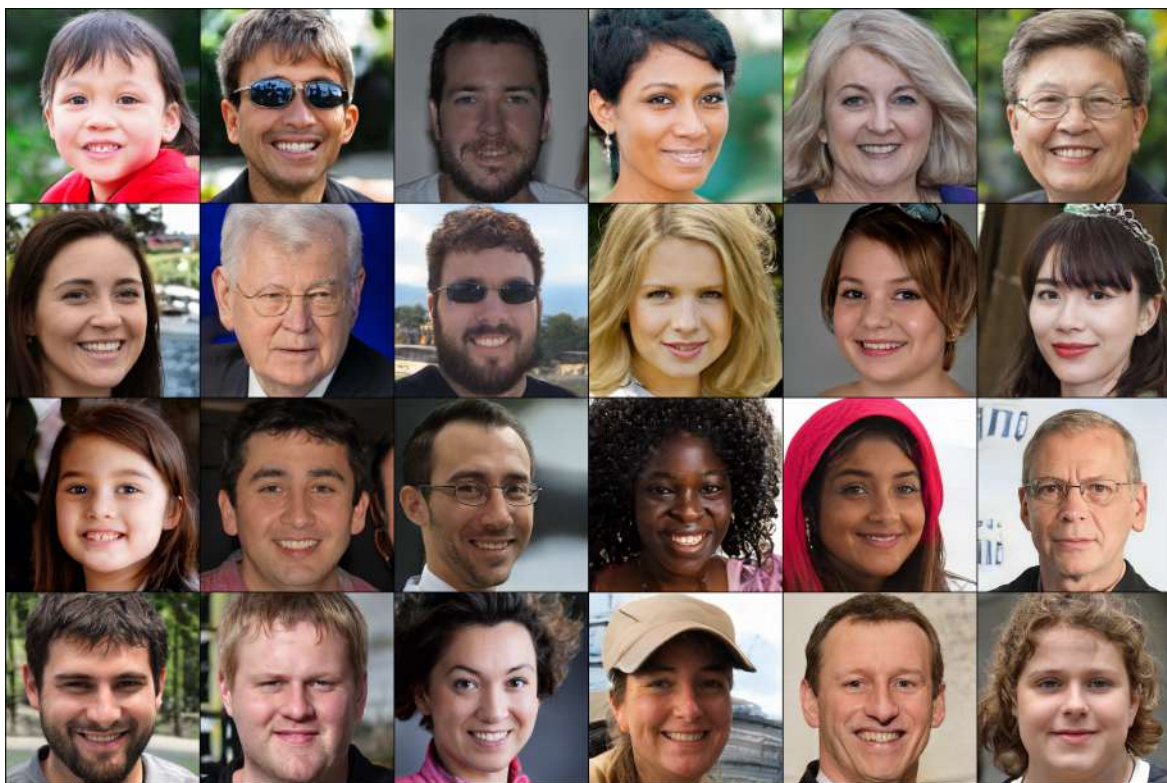


Figure 14. Images generated with **FQ**-StyleGAN2 on FFHQ-1024<sup>2</sup>.



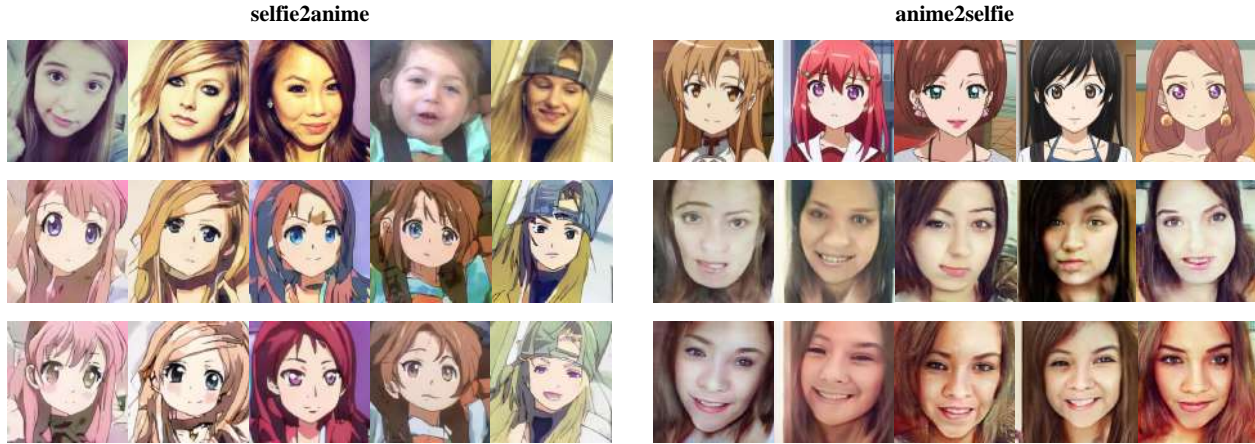


Figure 15. Visual comparisons on selfie2anime and anime2selfie. **First row:** input images. **Second row:** images generate by U-GAT-IT. **Third row:** images generated by FQ-U-GAT-IT.

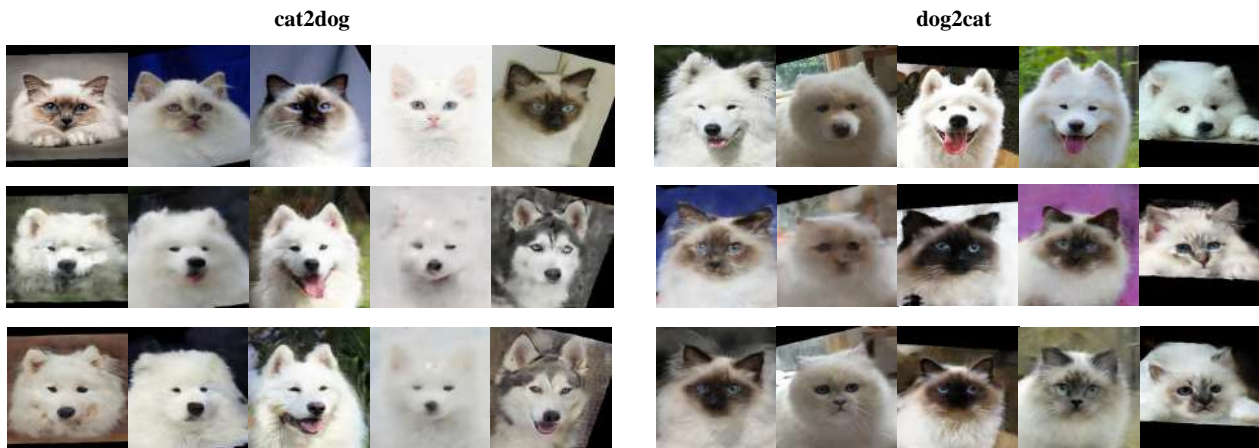


Figure 16. Visual comparisons on cat2dog and dog2cat. **First row:** input images. **Second row:** images generated by U-GAT-IT. **Third row:** images generated by FQ-U-GAT-IT.



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.

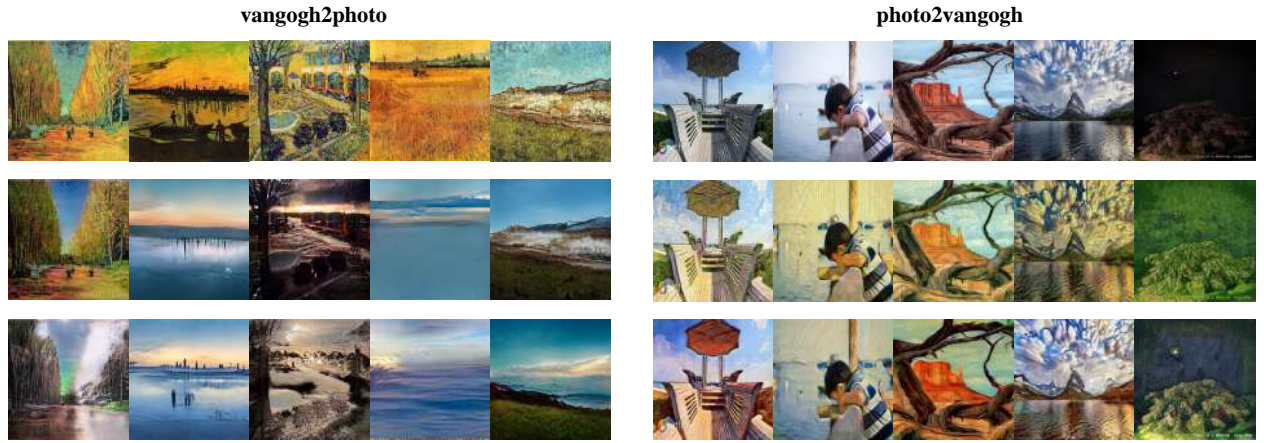


Figure 18. Visual comparisons on vangogh2photo and photo2vangogh. **First row:** input images. **Second row:** images generated by U-GAT-IT. **Third row:** images generated by FQ-U-GAT-IT.



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.









### Selfie2Anime (Translate from selfie to animation)


#### Scoring Standard

The goal of image translation is to blend a source domain and a target domain together so the output looks like the source domain, but "painted" in the style of the style target domain. So, a good translated image should contain the detail (e.g. texture and object) of the source image and style of the target domain.



#### Target domain samples



#### Test source image



#### Translated images



Which one do you prefer? Type in "1" if you vote for the left image, type in "2" if you like the right

Please choose your favorite translated image from the above two

Thanks for spending time >>

Submit

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

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

$z \in \mathbb{R}^{120} \sim \mathcal{N}(0, 1)$
$\text{Embed}(y) \in \mathbb{R}^{128}$
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, $3 \times 3$ Conv $ch \rightarrow 3$
Tanh
(a) <b>Generator</b>

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

Table 9. Discriminator architecture in StyleGAN and StyleGAN2

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

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

$z \in \mathbb{R}^{120} \sim \mathcal{N}(0, 1)$
$\text{Embed}(y) \in \mathbb{R}^{128}$
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 $(64 \times 64)$
ResBlock up $2ch \rightarrow ch$
BN, ReLU, $3 \times 3$ Conv $ch \rightarrow 3$
Tanh
(a) <b>Generator</b>

RGB image $x \in \mathbb{R}^{128 \times 128 \times 3}$
ResBlock up $ch \rightarrow 2ch$
Non-Local Block $(64 \times 64)$
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
$\text{Embed}(y)\mathbf{h} + (\text{linear} \rightarrow 1)$
(b) <b>Discriminator</b>

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