Alleviation of Gradient Exploding in GANs: Fake Can Be Real Supplementary Material

Song Tao, Jia Wang*
Department of Electronic Engineering
Shanghai Jiao Tong University
{taosong, jiawang}@sjtu.edu.cn

A. Proof for Proposition 1

For empirical discriminator, it maximizes the following objective:

$$\mathcal{L} = \mathbb{E}_{x\in D_r}[\log(D(x))] + \mathbb{E}_{y\in D_g}[\log(1-D(y))]. \quad (1)$$

Assume that samples are normalized:

$$||x_i|| = ||y_i|| = 1, \forall x \in D_r, y \in D_g. \quad (2)$$

Let $W_1 \in \mathbb{R}^{2\times d_x}$, $W_2 \in \mathbb{R}^{2\times 2}$ and $W_3 \in \mathbb{R}^2$ be the weight matrices, $b \in \mathbb{R}^2$ offset vector and k_1,k_2 a constant, We can construct needed discriminator as a MLP with two hidden layer containing $\mathcal{O}(2dim(x))$ parameters. We set weight matrices

$$W_1 = \begin{bmatrix} {x_0}^T \\ {y_0}^T \end{bmatrix}, W_2 = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}, W_3 = \begin{bmatrix} \frac{1}{2}+\frac{\epsilon}{2} \\ \frac{1}{2}-\frac{\epsilon}{2} \end{bmatrix}. \quad (3)$$

For any input $v \in D_r \cup D_g$, the discriminator output is computed as:

$$D(v) = {W_3}^T \sigma(k_2 W_2 \sigma(k_1(W_1 v - b))), \quad (4)$$

where $\sigma(x) = \frac{1}{1+e^{-x}}$ is the sigmoid function. Let $\alpha = W_1 v - b$, we have

$$\alpha_1 = \begin{cases} 1-b_1, \text{if } v = x_0 \\ l-b_1, \text{if } v \neq x_0 \end{cases}, \alpha_2 = \begin{cases} 1-b_2, \text{if } v = y_0 \\ l-b_2, \text{if } v \neq y_0 \end{cases}, \quad (5)$$

where $l < 1$. Let $\beta = \sigma(k_1\alpha)$, we have

$$\beta_1 = \begin{cases} 1, \text{if } v = x_0 \\ 0, \text{if } v \neq x_0 \end{cases}, \beta_2 = \begin{cases} 1, \text{if } v = y_0 \\ 0, \text{if } v \neq y_0 \end{cases} \quad (6)$$

as $k_1 \to \infty$ and $b \to 1^-$. Let $\gamma = \sigma(k_2 W_2 \beta)$, we have

$$\gamma_1 = \begin{cases} 1, & \text{if } v = x_0 \\ 0, & \text{if } v = y_0 \\ \frac{1}{2}, & \text{if } v \neq x_0, y_0 \end{cases}, \gamma_2 = \begin{cases} 0, & \text{if } v = x_0 \\ 1, & \text{if } v = y_0 \\ \frac{1}{2}, & \text{if } v \neq x_0, y_0 \end{cases} \quad (7)$$

as $k_2 \to \infty$. Hence, for any input $v \in D_r \cup D_g$, discriminator outputs

$$D(v) = {W_3}^T \gamma = \begin{cases} \frac{1}{2}+\frac{\epsilon}{2}, \text{if } v = x_0 \\ \frac{1}{2}-\frac{\epsilon}{2}, \text{if } v = y_0 \\ \frac{1}{2}, \text{else} \end{cases}. \quad (8)$$

In this case, the discriminator objective has a more optimal value than the theoretical optimal version:

$$\begin{aligned} \mathcal{L} &= \frac{1}{n}((n-1)\log\frac{1}{2} + \log(\frac{1}{2}+\frac{\epsilon}{2})) \\ &\quad + \frac{1}{m}((m-1)\log\frac{1}{2} + \log(\frac{1}{2}+\frac{\epsilon}{2})) \\ &> 2\log\frac{1}{2}. \end{aligned} \quad (9)$$

Then the discriminator outputs a constant $\frac{1}{2}$ except that $D(x_0) = \frac{1}{2}+\frac{\epsilon}{2}$ and $D(y_0) = \frac{1}{2}-\frac{\epsilon}{2}$ satisfying the condition.

B. Proof for proposition 2

We rewrite $f(\xi_0, \xi_1, \cdots, \xi_{m_0})$ here

$$f = \log\sigma(\xi_0) + \frac{n}{m}\sum_{i=1}^{m_0}\log(1-\sigma(\xi_i)) - \frac{nk}{m_0}\sum_{i=1}^{m_0}(\xi_0-\xi_i)^2. \quad (10)$$

To achieve the optimal value, let $f'(\xi_i) = 0, i = 0, \cdots, m_0$ and we have

$$f'(\xi_0^*) = 1 - \sigma(\xi_0^*) - \frac{2nk}{m_0}\sum_{i=1}^{m_0}(\xi_0^* - \xi_i^*) = 0, \quad (11)$$

$$f'(\xi_i^*) = -\frac{n}{m}\sigma(\xi_i^*) + \frac{2nk}{m_0}(\xi_0^* - \xi_i^*) = 0, i = 1, \cdots, m_0. \quad (12)$$

*Corresponding author

It is obvious that $\xi_1^* = \xi_2^* = \cdots = \xi_{m_0}^* = \xi^*$. Hence we have

$$1 - \sigma(\xi_0^*) - 2nk(\xi_0^* - \xi^*) = 0, \tag{13}$$

$$-\frac{n}{m}\sigma(\xi^*) + \frac{2nk}{m_0}(\xi_0^* - \xi^*) = 0. \tag{14}$$

We can solve

$$\xi^* = -\ln(\frac{nm_0}{m\sigma(-\xi_0^*)} - 1). \tag{15}$$

Substitute Eqn. 15 into Eqn. 13 and we get

$$f'(\xi_0^*) = \sigma(-\xi_0^*) - 2nk(\xi_0^* + \ln(\frac{nm_0}{m\sigma(-\xi_0^*)} - 1)) = 0. \tag{16}$$

We can also have from Eqn. 15 and Eqn. 13 respectively

$$\xi_0^* - \xi^* = \xi_0^* + \ln(\frac{nm_0}{m\sigma(-\xi_0^*)} - 1), \tag{17}$$

$$= \frac{\sigma(-\xi_0^*)}{2nk}. \tag{18}$$

To satisfy Eqn.16, $\xi_0^* + \log(\frac{nm_0}{m\sigma(-\xi_0^*)} - 1) > 0$. Hence, with k increasing, ξ_0^* decreases from Eqn.16. Based on Eqn.17, we further know with k increasing, ξ^* increases, $\xi_0^* - \xi^*$ decreases and $\sigma(-\xi_i^*)(\xi_0^* - \xi_i^*)$ decreases. Similarly, based on Eqn.18 and Eqn.16, we can achieve with m_0 increasing, $\sigma(-\xi_i^*)(\xi_0^* - \xi_i^*)$ increases finishing the proof.

C. Proof for proposition 3

Similar to the proof for Proposition 2, let $h'(\xi_0) = h'(\xi_i) = 0, i = 1, \cdots, m_0$, and we can easily achieve $\xi_1^* = \xi_2^* = \cdots = \xi_{m_0}^*$,

$$\lambda(\xi_0^*) = \frac{\sigma(-\xi_0^*)}{nm_0}[e^{2nk\xi_0^* - \sigma(-\xi_0^*)}(\frac{nm_0}{m\sigma(-\xi_0^*)} - 1)^{2nk} - 1], \tag{19}$$

and

$$\xi_0^* - \xi_i^* = \frac{\sigma(-\xi_0^*)}{2nk}. \tag{20}$$

It can be easily proved that $\lambda'(\xi_0^*) > 0$. To satisfy Eqn.19, with λ increasing, ξ_0^* increases, and, when $\lambda \to \infty$, $\xi_0^* \to \infty$. Based on Eqn.20, we further know with λ increasing, ξ_i^* increases, $\xi_0^* - \xi_i^*$ decreases, and $\sigma(-\xi_i^*)(\xi_0^* - \xi_i^*)$ decreases. We can also achieve that when $\lambda \to \infty$, $\sigma(-\xi_i^*)(\xi_0^* - \xi_i^*) \to 0$, finishing the proof.

D. Network architectures

We give the network architectures used in our experiments and a simple DCGAN based architecture is also included in CIFAR experiment.

Table 1. Generator architecture in synthetic experiment

Layer	output size	filter
Fully connected	64	$2 \to 64$
RELU	64	-
Fully connected	64	$64 \to 64$
RELU	64	-
Fully connected	64	$64 \to 64$
RELU	64	-
Fully connected	2	$64 \to 2$

Table 2. Discriminator architecture in synthetic experiment

Layer	output size	filter
Fully connected	64	$2 \to 64$
RELU	64	-
Fully connected	64	$64 \to 64$
RELU	64	-
Fully connected	64	$64 \to 64$
RELU	64	-
Fully connected	1	$64 \to 1$

Table 3. Generator DCGAN based architecture in CIFAR experiment

Layer	output size	filter
Fully connected	$256 \cdot 4 \cdot 4$	$128 \to 256 \cdot 4 \cdot 4$
Reshape	$256 \times 4 \times 4$	-
TransposedConv2D	$128 \times 8 \times 8$	$256 \to 128$
TransposedConv2D	$64 \times 16 \times 16$	$128 \to 64$
TransposedConv2D	$3 \times 32 \times 32$	$64 \to 3$

Table 4. Discriminator DCGAN based architecture in CIFAR experiment

Layer	output size	filter
Conv2D	$64 \times 16 \times 16$	$3 \to 64$
Conv2D	$128 \times 8 \times 8$	$64 \to 128$
Conv2D	$256 \times 4 \times 4$	$128 \to 256$
Reshape	$256 \cdot 4 \cdot 4$	-
Fully Connected	$256 \cdot 4 \cdot 4$	$256 \cdot 4 \cdot 4 \to 1$

Table 5. Generator ResNet architecture in CIFAR experiment

Layer	output size	filter
Fully connected	$512 \cdot 4 \cdot 4$	$128 \rightarrow 512 \cdot 4 \cdot 4$
Reshape	$512 \times 4 \times 4$	-
Resnet-Block	$256 \times 4 \times 4$	$512 \rightarrow 256 \rightarrow 256$
NN-Upsampling	$256 \times 8 \times 8$	-
Resnet-Block	$128 \times 8 \times 8$	$256 \rightarrow 128 \rightarrow 128$
NN-Upsampling	$128 \times 16 \times 16$	-
Resnet-Block	$64 \times 16 \times 16$	$128 \rightarrow 64 \rightarrow 64$
NN-Upsampling	$64 \times 32 \times 32$	-
Resnet-Block	$64 \times 32 \times 32$	$64 \rightarrow 64 \rightarrow 64$
Conv2D	$3 \times 32 \times 32$	$64 \rightarrow 3$

Table 6. Discriminator ResNet architecture in CIFAR experiment

Layer	output size	filter
Conv2D	$64 \times 32 \times 32$	$3 \rightarrow 64$
Resnet-Block	$128 \times 32 \times 32$	$64 \rightarrow 64 \rightarrow 128$
Avg-Pool2D	$128 \times 16 \times 16$	-
Resnet-Block	$256 \times 16 \times 16$	$128 \rightarrow 128 \rightarrow 256$
Avg-Pool2D	$256 \times 8 \times 8$	-
Resnet-Block	$512 \times 8 \times 8$	$256 \rightarrow 256 \rightarrow 512$
Avg-Pool2D	$512 \times 4 \times 4$	-
Reshape	$512 \cdot 4 \cdot 4$	-
Fully Connected	1	$512 \cdot 4 \cdot 4 \rightarrow 1$

Table 7. Generator conventional architecture in CIFAR experiment

Layer	output size	filter
Fully connected	$512 \cdot 1 \cdot 1$	$128 \rightarrow 512 \cdot 1 \cdot 1$
Reshape	$512 \times 1 \times 1$	-
TransposedConv2D	$512 \times 4 \times 4$	$512 \rightarrow 512$
TransposedConv2D	$512 \times 4 \times 4$	$512 \rightarrow 512$
NN-Upsampling	$512 \times 8 \times 8$	-
TransposedConv2D	$256 \times 8 \times 8$	$512 \rightarrow 256$
TransposedConv2D	$256 \times 8 \times 8$	$256 \rightarrow 256$
NN-Upsampling	$256 \times 16 \times 16$	-
TransposedConv2D	$128 \times 16 \times 16$	$256 \rightarrow 128$
TransposedConv2D	$128 \times 16 \times 16$	$128 \rightarrow 128$
NN-Upsampling	$128 \times 32 \times 32$	-
TransposedConv2D	$64 \times 32 \times 32$	$128 \rightarrow 64$
TransposedConv2D	$64 \times 32 \times 32$	$64 \rightarrow 64$
TransposedConv2D	$3 \times 32 \times 32$	$64 \rightarrow 3$

Table 8. Discriminator conventional architecture in CIFAR experiment

Layer	output size	filter
Conv2D	$64 \times 32 \times 32$	$3 \rightarrow 64$
Conv2D	$64 \times 32 \times 32$	$64 \rightarrow 64$
Conv2D	$128 \times 32 \times 32$	$64 \rightarrow 128$
Avg-Pool2D	$128 \times 16 \times 16$	-
Conv2D	$128 \times 16 \times 16$	$128 \rightarrow 128$
Conv2D	$256 \times 16 \times 16$	$128 \rightarrow 256$
Avg-Pool2D	$256 \times 8 \times 8$	-
Conv2D	$256 \times 8 \times 8$	$256 \rightarrow 256$
Conv2D	$512 \times 8 \times 8$	$256 \rightarrow 512$
Avg-Pool2D	$512 \times 4 \times 4$	-
Conv2D	$512 \times 4 \times 4$	$512 \rightarrow 512$
Conv2D	$512 \times 4 \times 4$	$512 \rightarrow 512$
Reshape	$512 \cdot 4 \cdot 4$	-
Fully Connected	1	$512 \cdot 4 \cdot 4 \rightarrow 1$

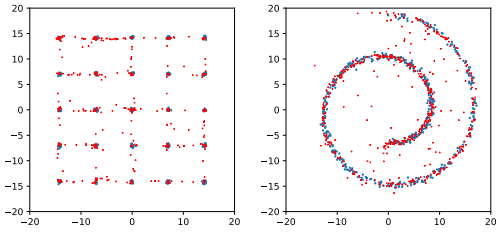
Table 9. Generator architecture in ImageNet experiment

Layer	output size	filter
Fully connected	$1024 \cdot 4 \cdot 4$	$256 \rightarrow 1024 \cdot 4 \cdot 4$
Reshape	$1024 \times 4 \times 4$	-
Resnet-Block	$1024 \times 4 \times 4$	$1024 \rightarrow 1024 \rightarrow 1024$
Resnet-Block	$1024 \times 4 \times 4$	$1024 \rightarrow 1024 \rightarrow 1024$
NN-Upsampling	$1024 \times 8 \times 8$	-
Resnet-Block	$512 \times 8 \times 8$	$1024 \rightarrow 512 \rightarrow 512$
Resnet-Block	$512 \times 8 \times 8$	$512 \rightarrow 512 \rightarrow 512$
NN-Upsampling	$512 \times 16 \times 16$	-
Resnet-Block	$256 \times 16 \times 16$	$512 \rightarrow 256 \rightarrow 256$
Resnet-Block	$256 \times 16 \times 16$	$256 \rightarrow 256 \rightarrow 256$
NN-Upsampling	$256 \times 32 \times 32$	-
Resnet-Block	$128 \times 32 \times 32$	$256 \rightarrow 128 \rightarrow 128$
Resnet-Block	$128 \times 32 \times 32$	$128 \rightarrow 128 \rightarrow 128$
NN-Upsampling	$128 \times 64 \times 64$	-
Resnet-Block	$64 \times 64 \times 64$	$128 \rightarrow 64 \rightarrow 64$
Resnet-Block	$64 \times 64 \times 64$	$64 \rightarrow 64 \rightarrow 64$
Conv2D	$3 \times 64 \times 64$	$64 \rightarrow 3$

Table 10. Discriminator architecture in ImageNet experiment

Layer	output size	filter
Conv2D	$64 \times 64 \times 64$	$3 \rightarrow 64$
Resnet-Block	$64 \times 64 \times 64$	$64 \rightarrow 64 \rightarrow 64$
Resnet-Block	$128 \times 64 \times 64$	$64 \rightarrow 64 \rightarrow 128$
Avg-Pool2D	$128 \times 32 \times 32$	-
Resnet-Block	$128 \times 32 \times 32$	$128 \rightarrow 128 \rightarrow 128$
Resnet-Block	$256 \times 32 \times 32$	$128 \rightarrow 128 \rightarrow 256$
Avg-Pool2D	$256 \times 16 \times 16$	-
Resnet-Block	$256 \times 16 \times 16$	$256 \rightarrow 256 \rightarrow 256$
Resnet-Block	$512 \times 16 \times 16$	$256 \rightarrow 256 \rightarrow 512$
Avg-Pool2D	$512 \times 8 \times 8$	-
Resnet-Block	$512 \times 8 \times 8$	$512 \rightarrow 512 \rightarrow 512$
Resnet-Block	$1024 \times 8 \times 8$	$512 \rightarrow 512 \rightarrow 1024$
Avg-Pool2D	$1024 \times 4 \times 4$	-
Resnet-Block	$1024 \times 4 \times 4$	$1024 \rightarrow 1024 \rightarrow 1024$
Resnet-Block	$1024 \times 4 \times 4$	$1024 \rightarrow 1024 \rightarrow 1024$
Fully Connected	1	$1024 \cdot 4 \cdot 4 \rightarrow 1$

E. Further results



(a) (b)

Figure 1. Generation of our method on a mixture of 25 Gaussians dataset and swissroll datatset.

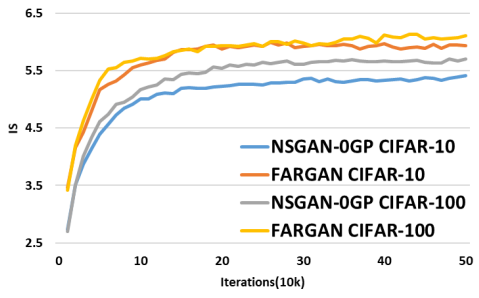
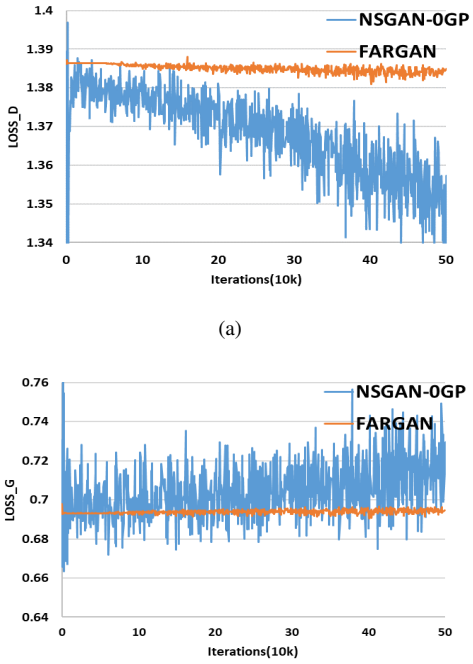
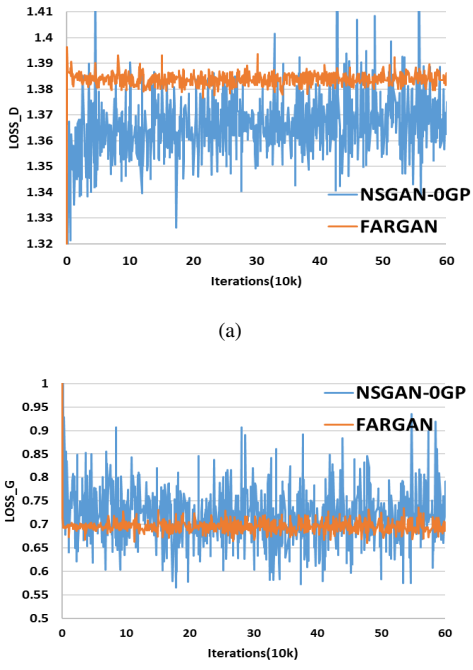


Figure 2. Inception score on CIFAR-10 and CIFAR-100 of NSGAN-0GP and FARGAN for a DCGAN based network architecture. Our method still outperforms NSGAN-0GP.



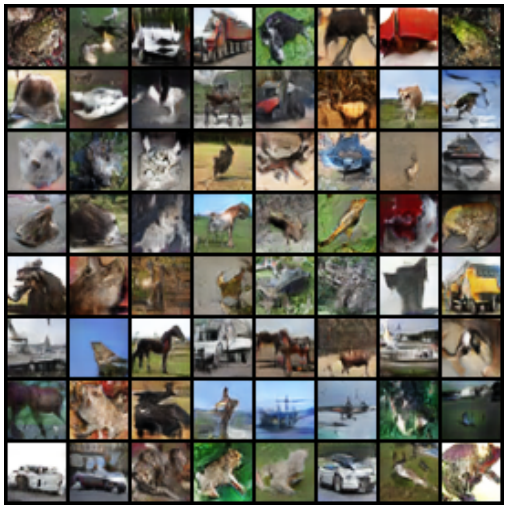
(a) (b)

Figure 3. Losses of discriminator (not including regularization term) and generator on CIFAR-100 of NSGAN-0GP and FARGAN

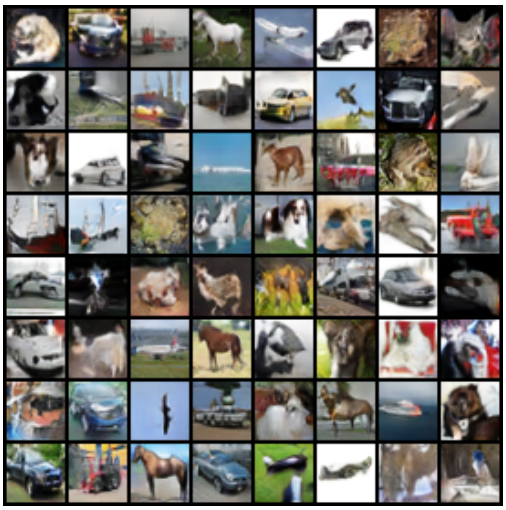


(a) (b)

Figure 4. Losses of discriminator (not including regularization term) and generator on ImageNet of NSGAN-0GP and FARGAN



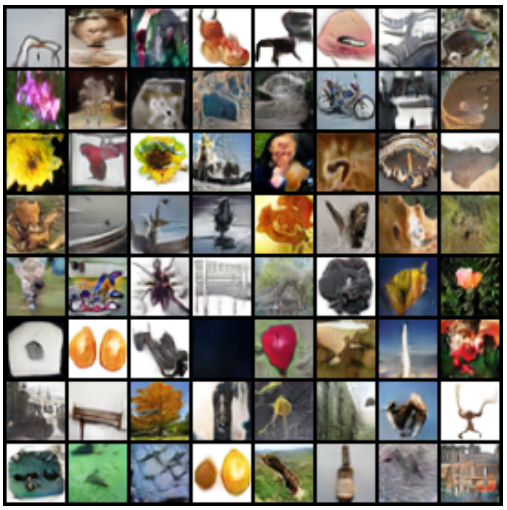
(a) image generation of NSGAN-0GP



(b) image generation of FARGAN

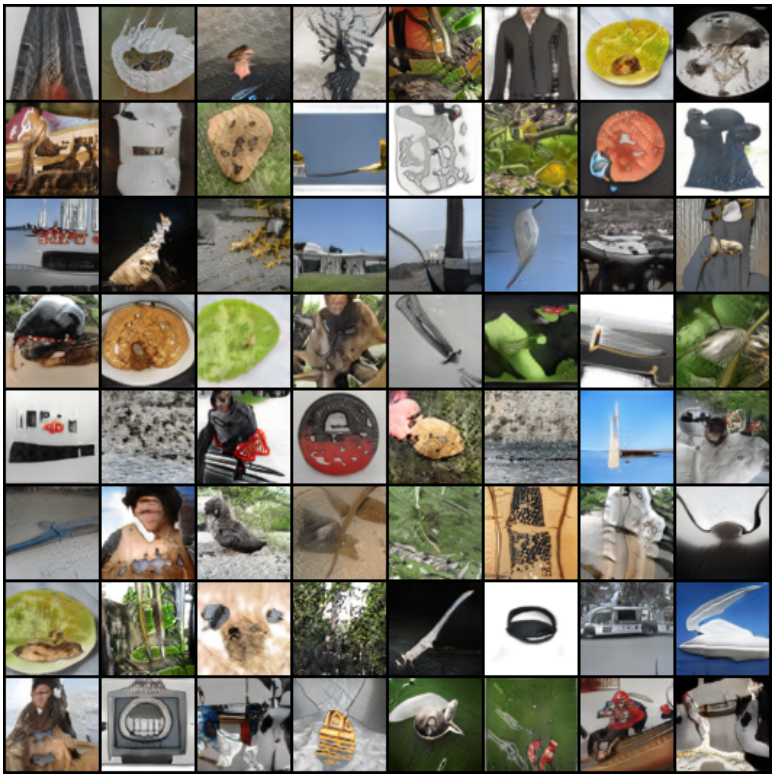
Figure 5. Image generation of CIFAR-10.

(a) image generation of NSGAN-0GP

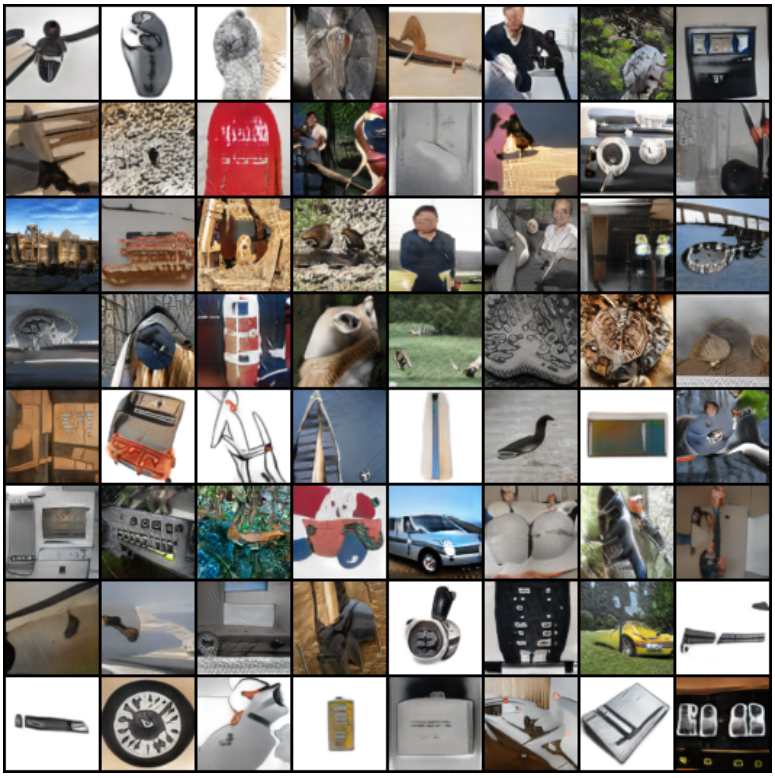


(b) image generation of FARGAN

Figure 6. Image generation of CIFAR-100.



(a) image generation of NSGAN-0GP



(b) image generation of FARGAN

Figure 7. Image generation of ImageNet.