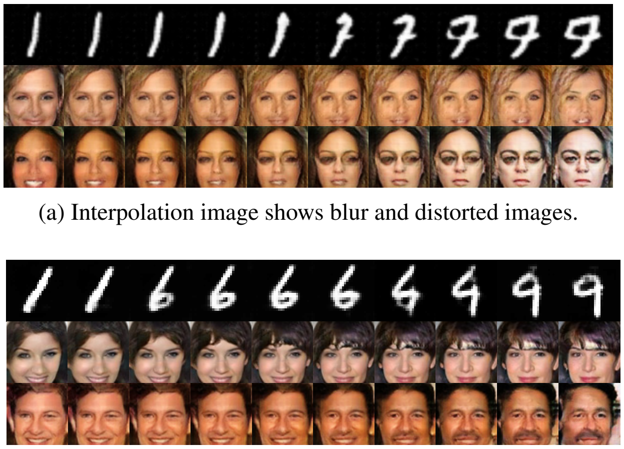
The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)

Smooth Deep Image Generator from Noises

Tianyu Guo,[1,2,3] **Chang Xu,**[2] **Boxin Shi,**[4] **Chao Xu,**[1,3] **Dacheng Tao**[2]
[1]Key Laboratory of Machine Perception (MOE), School of EECS, Peking University, China
[2]UBTECH Sydney AI Centre, School of Computer Science, FEIT, University of Sydney, Australia
[3]Cooperative Medianet Innovation Center, Peking University, China
[4]National Engineering Laboratory for Video Technology, School of EECS, Peking University, China
tianyuguo@pku.edu.cn, c.xu@sydney.edu.au, shiboxin@pku.edu.cn,
xuchao@cis.pku.edu.cn, dacheng.tao@sydney.edu.au

Abstract

Generative Adversarial Networks (GANs) have demonstrated a strong ability to fit complex distributions since they were presented, especially in the field of generating natural images. Linear interpolation in the noise space produces a continuously changing in the image space, which is an impressive property of GANs. However, there is no special consideration on this property in the objective function of GANs or its derived models. This paper analyzes the perturbation on the input of the generator and its influence on the generated images. A smooth generator is then developed by investigating the tolerable input perturbation. We further integrate this smooth generator with a gradient penalized discriminator, and design smooth GAN that generates stable and high-quality images. Experiments on real-world image datasets demonstrate the necessity of studying smooth generator and the effectiveness of the proposed algorithm.



(a) Interpolation image shows blur and distorted images.

(b) Smooth interpolation shows clear and high-quality images.

Figure 1: Interpolation images generated from WGAN-GP (Gulrajani et al. 2017) (a) and the proposed smooth GAN (b).

Introduction

Deep generative models have attracted increasing attention from researchers, especially in the task of natural image generation. Representative techniques include Variational Auto-Encoder (VAE) (Kingma and Welling 2013), PixelCNN (van den Oord et al. 2016), and Generative Adversarial Networks (GANs) (Goodfellow et al. 2014). Generative Adversarial Networks (GANs) (Goodfellow et al. 2014) translate Gaussian inputs into natural images by discovering the equilibrium within a max-min game. The generator in vanilla GANs is to transform noisy vectors into images, while the discriminator aims to distinguish the generated samples from real samples. Convincing images generated from noisy vectors through GANs could be employed to augment image datasets, which would alleviate the shortage of training data in some tasks. Moreover, image-to-image translation (Chen et al. 2018; 2019) based on GANs also gets its popularity.

However, vanilla GANs have flaws in its stability, and we have seen many promising works to alleviate this problem by modifying the network frameworks or proposing improved loss functions (Radford, Metz, and Chintala 2015; Nguyen et al. 2017; Karras et al. 2017; Mao et al. 2017; Berthelot, Schumm, and Metz 2017a; Arjovsky, Chintala, and Bottou 2017; Gulrajani et al. 2017). Besides, a considerable body of work has been conducted to arbitrarily manipulate generated images according to different factors, *e.g.*, the category, illumination, and style (Chen et al. 2016). Beyond meaningless noise input in GANs, interpretable features can be discovered by investigating label information in conditional GANs (Mirza and Osindero 2014), exploring the mutual information between elements of input in infoGAN (Chen et al. 2016) or leveraging the discriminator on latent space in AAE (Makhzani et al. 2015).

Noise vector inputs for GANs can be taken as low-dimensional representations of images. As widely accepted in representation learning, the closeness of two data points is supposed to be preserved before and after transformation. Most of these improved GANs methods implicitly assume that the generator would translate linear interpolation in the input noise space to semantic interpolation in the output image space (Bojanowski et al. 2017). Although this kind of experimental result showing interesting visual effects attracts readers' attention, the quality of images generated through interpolations could be very noisy and fragile, and some of these images would look obviously unnatural or even meaningless, as demonstrated in Figure 1(a). Efforts are spent towards generating high-quality images or stabilizing the training of GANs, and how to ensure the success of semantic interpolation in GANs has rarely been investigated.

In this paper, we propose a smooth deep image generator that can suppress the influence of input perturbations on generated images. By investigating the connections between input noises and generated images, we theoretically present the most serious input perturbation that can be tolerated for an output image of desired precision. A gradient-based loss function is then introduced to reduce the variation of generated images caused by perturbations on input noises, which encourages a smooth interpolation of images. Combining a discriminator with gradient penalty, we show the smooth generator will be beneficial for improving the quality of interpolation samples, as demonstrated in Figure 1(b). Experimental results on real-world datasets MNIST (LeCun et al. 1998), CIFAR-10 (Krizhevsky and Hinton 2009), and CelebA (Liu et al. 2015) demonstrate the generator produced by the proposed method is essential for the success of smooth and high-quality interpolation of images.

Related Work

In this section, we briefly introduce related works on generative adversarial networks (GANs).

Although the GANs model has powerful image generation capabilities, the model was often trapped in the problem of unstable training and difficulty in convergence. Some methods have been proposed to solve this problem. DCGAN (Radford, Metz, and Chintala 2015) introduced a network structure that works well and is stable. WGAN (Arjovsky, Chintala, and Bottou 2017) proved the defect of vanilla adversarial loss and proposed Wasserstein distance to measure the distance between the generated data distribution and the real data distribution. However, the weight clip used in WGAN to ensure the Lipschitz continuous of D leads to the loss of the capacity of neural networks. To solve this problem, WGAN-GP (Gulrajani et al. 2017) proposed gradient penalty instead of weight clip operation to satisfy Lipschitz continuous condition. BEGAN (Berthelot, Schumm, and Metz 2017a) proposed a novel concept of equilibrium that can help GANs to achieve considerable results using standard training methods that do not incorporate tricks. At the same time, similar to the Wasserstein distance, this degree of equilibrium can estimate the degree of convergence of the model. MMD GAN (Li et al. 2017b) connected moment matching network and GANs and achieved competitive performances with state-of-the art GANs. Kim et al. (Kim and Bengio 2016) and VGAN (Zhai et al. 2016) integrated GANs with the energy-based model and improved the performance of generative models.

GANs have achieved remarkable results in image generation. LapGAN (Denton et al. 2015) generated high-resolution images from low resolution one with the help of the Laplacian pyramid framework. Furthermore, ProgGAN (Karras et al. 2017) proposed to train generator and discriminator progressively at upscale resolution levels, which can produce extremely high-quality 2k resolution images. In semi-supervised learning, TripleGAN (Li et al. 2017a) introduced a classifier C to perform generation tasks under semi-supervised conditions. DCGAN (Radford, Metz, and Chintala 2015) introduced interpolation in latent space generate the smooth transition in image space. However, there is no insurance for the sign of smooth transition in the adversarial Loss. As a result, this paper analyzes the constraint required by the smooth transition in image space and introduces a method to enhance this sign of GANs.

Proposed Method

In this section, we analyze the conditions required by the smooth transition in image space and develop a smooth generator within GANs.

Generative Adversarial Nets

A discriminator D and a generator G play a max-min game in GANs, in which the discriminator D is responsible for distinguishing real samples from generated samples, while the generator G is to deceive the discriminator D. When the game achieves equilibrium, the generator G would be able to fit complicated distribution of real samples.

Formally given the sample $\boldsymbol{x}$ from the real distribution $\mathbb{P}_d$ and the noise $\boldsymbol{z}$ drawn from noise distribution $\mathbb{P}_z$ (*e.g.*, Gaussian or uniform distribution), the optimal generator G transforming the noise distribution to the real data distribution can be solved from the following min-max optimization problem:

$$\min_G \max_D \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathbb{P}_d}[\log(D(\boldsymbol{x}))] + \mathop{\mathbb{E}}_{\boldsymbol{z}\sim\mathbb{P}_z}[\log(D(G(\boldsymbol{z})))]. \quad (1)$$

We denote the distribution of generated sample $G(z)$ as $\mathbb{P}_G$. By alternately optimizing the generator G and the discriminator D in the min-max problem, we expect that the difference between the generated distribution $\mathbb{P}_G$ and the real data distribution $\mathbb{P}_d$ would be gradually consistent with each other.

Smooth Generator

In the noise-to-image generation task, it is difficult to know what type of perturbation could happen in practice. Hence we consider the general perturbation on pixels, and Euclidean distance is adopted for the measurement. Considering a continuous translation or rotation, generated images are still expected to evolve smoothly, and thus pixel values should avoid sudden changes. Given the input noise vector $\boldsymbol{z}$ and the generator G, the generated image can be written as $G(\boldsymbol{z})$. We suppose that the value of the i-th pixel on the image $G(\boldsymbol{z})$ is determined by $G_i(\boldsymbol{z})$, where $G_i(\cdot)$ is reduced from $G(\cdot)$. A smooth generator is then expected to have the following pixel-wise property:

$$\big|G_i(\boldsymbol{z}+\boldsymbol{\delta}) - G_i(\boldsymbol{z})\big| < \epsilon \quad (2)$$

where $\boldsymbol{\delta}$ stands for a small perturbation over input noise $\boldsymbol{z}$, and $\epsilon > 0$ is a small constant number. Since linear interpolation around $\boldsymbol{z}$ in the noise space can be approximated as imposing perturbation $\boldsymbol{\delta}$ on $\boldsymbol{z}$, Eq. (2) would encourage the image generated from noise interpolation would not be far from the original image. In addition, Eq. (2) can be helpful to improve the robustness of the generator, so that it would not be easily disabled by adversarial noise inputs with slight perturbations. However, it is difficult to straightforwardly integrate Eq. (2) into objective function of GANs, because of the unspecified $\boldsymbol{\delta}$. We next proceed to analyze the appropriate $\boldsymbol{\delta}$ that satisfies Eq. (2) in the following theorem.

Theorem 1. *Fix $\epsilon > 0$. Given $\boldsymbol{z} \in \mathbb{R}^d$ as the noise input of generator G_i, if the perturbation $\boldsymbol{\delta} \in \mathbb{R}^d$ satisfies*

$$\|\boldsymbol{\delta}\|_q < \frac{\epsilon}{\max_{\hat{\boldsymbol{z}} \in B_p(\boldsymbol{z},R)} \|\nabla_{\hat{\boldsymbol{z}}} G_i(\hat{\boldsymbol{z}})\|_p}, \tag{3}$$

with $\frac{1}{p} + \frac{1}{q} = 1$, we have $\left|G_i(\boldsymbol{z} + \boldsymbol{\delta}) - G_i(\boldsymbol{z})\right| < \epsilon$.

Proof. Without loss of generality, we first suppose $G_i(\boldsymbol{z}) \leq G_i(\boldsymbol{z} + \boldsymbol{\delta})$. Our aim is then to demonstrate what condition $\boldsymbol{\delta}$ should obey to realize

$$0 \leq G_i(\boldsymbol{z} + \boldsymbol{\delta}) - G_i(\boldsymbol{z}) < \epsilon. \tag{4}$$

By the main theorem of calculus, we have

$$G_i(\boldsymbol{z} + \boldsymbol{\delta}) = G_i(\boldsymbol{z}) + \int_0^1 \langle \nabla_{\boldsymbol{z}} G_i(\boldsymbol{z} + t\boldsymbol{\delta}), \boldsymbol{\delta} \rangle \mathrm{d}t, \tag{5}$$

so that

$$0 \leq G_i(\boldsymbol{z} + \boldsymbol{\delta}) - G_i(\boldsymbol{z}) = \int_0^1 \langle \nabla_{\boldsymbol{z}} G_i(\boldsymbol{z} + t\boldsymbol{\delta}), \boldsymbol{\delta} \rangle \mathrm{d}t. \tag{6}$$

Consider the fact that

$$\begin{aligned} &\int_0^1 \langle \nabla_{\boldsymbol{z}} G_i(\boldsymbol{z} + t\boldsymbol{\delta}), \boldsymbol{\delta} \rangle \mathrm{d}t \\ &\qquad \leq \|\boldsymbol{\delta}\|_q \int_0^1 \|\nabla_{\boldsymbol{z}} G_i(\boldsymbol{z} + t\boldsymbol{\delta})\|_p \mathrm{d}t, \end{aligned} \tag{7}$$

where holder inequality is applied and q-norm is dual to the p-norm with $\frac{1}{p} + \frac{1}{q} = 1$. Suppose that $\hat{\boldsymbol{z}} = \boldsymbol{z} + t\boldsymbol{\delta}$ lies in a sphere centered at $\boldsymbol{z}$ with a radius R, and we define the sphere as $B_p(\boldsymbol{z}, R) = \{\hat{\boldsymbol{z}} \in \mathbb{R}^d \mid \|\boldsymbol{z} - \hat{\boldsymbol{z}}\|_p \leq R\}$. Hence, we have

$$\int_0^1 \|\nabla_{\boldsymbol{z}} G(\boldsymbol{z} + t\boldsymbol{\delta})\|_p \mathrm{d}t \leq \max_{\hat{\boldsymbol{z}} \in B_p(\boldsymbol{z},R)} \|\nabla_{\hat{\boldsymbol{z}}} G_i(\hat{\boldsymbol{z}})\|_p. \tag{8}$$

By combining Eqs. (7) and (8), Eq. (6) can be re-written as

$$0 \leq G_i(\boldsymbol{z} + \boldsymbol{\delta}) - G_i(\boldsymbol{z}) \leq \|\boldsymbol{\delta}\|_q \max_{\hat{\boldsymbol{z}} \in B_p(\boldsymbol{z},R)} \|\nabla_{\hat{\boldsymbol{z}}} G_i(\hat{\boldsymbol{z}})\|_p. \tag{9}$$

If the right side of Eq. (9) is always upper bounded by ϵ, *i.e.*,

$$\|\boldsymbol{\delta}\|_q \max_{\hat{\boldsymbol{z}} \in B_p(\boldsymbol{z},R)} \|\nabla_{\hat{\boldsymbol{z}}} G_i(\hat{\boldsymbol{z}})\|_p < \epsilon, \tag{10}$$

we can achieve the conclusion that $0 < G_i(z + \delta) - G_i(z) < \epsilon$. According to Eq. (10), $\boldsymbol{\delta}$ should satisfy

$$\|\boldsymbol{\delta}\|_q < \frac{\epsilon}{\max_{\hat{\boldsymbol{z}} \in B_p(\boldsymbol{z},R)} \|\nabla_{\hat{\boldsymbol{z}}} G_i(\hat{\boldsymbol{z}})\|_p}. \tag{11}$$

By setting $\boldsymbol{z} := \boldsymbol{z} + \boldsymbol{\delta}$ and $\boldsymbol{\delta} := -\boldsymbol{\delta}$, we we can get the same constraint (*i.e.*, Eq. (11)) over $\boldsymbol{\delta}$ to achieve $-\epsilon < G_i(\boldsymbol{z} + \boldsymbol{\delta}) - G_i(\boldsymbol{z})$. The proof is completed. □

By minimizing the denominator in Eq. (3), the model is expected to tolerate larger perturbation $\boldsymbol{\delta}$ under fixed difference ϵ on the i-th pixel. If all pixels of the generated image are simultaneously investigated, we then have

$$\mathcal{L} = \min_G \max_{\hat{z} \in B_p(z,R)} \|\nabla_{\hat{z}} G(\hat{z})\|_p. \tag{12}$$

However, $\max_{\hat{\boldsymbol{z}} \in B_p(\boldsymbol{z},R)} \|\nabla_{\hat{\boldsymbol{z}}} G(\hat{\boldsymbol{z}})\|_p$ is difficult to calculate. Since $\hat{\boldsymbol{z}}$ lies in a local region around $\boldsymbol{z}$, it is reasonable to assume that there is a data point $\hat{\boldsymbol{z}} \sim \mathbb{P}_z$ that can well approximate $\hat{\boldsymbol{z}}$. Hence, we can reformulate Eq. (12) as

$$\mathcal{L} = \min_G \mathop{\mathbb{E}}_{\boldsymbol{z} \sim \mathbb{P}_z} \|\nabla_{\boldsymbol{z}} G(\boldsymbol{z})\|_p. \tag{13}$$

Though minimizing $\|\nabla_{\boldsymbol{z}} G(\boldsymbol{z})\|_p$ will increase the perturbation $\boldsymbol{\delta}$ that can be tolerated by the generator, it is inappropriate to expect an enormously large value of $\boldsymbol{\delta}$, which could damage the diversity of generated images. If the generator is extremely insensitive to changes in the input, linear interpolation in noise space would always lead to the same output. As a result, we introduce a constant number k as a margin to constrain the value of $\|\nabla_{\boldsymbol{z}} G(\boldsymbol{z})\|_p$,

$$\mathcal{L} = \mathop{\mathbb{E}}_{\boldsymbol{z} \sim \mathbb{P}_z} \max(0, \|\nabla_{\boldsymbol{z}} G(\boldsymbol{z})\|_p^2 - k). \tag{14}$$

If the value of $\|\nabla_{\boldsymbol{z}} G(\boldsymbol{z})\|_p$ is larger than k, there will be penalty on the generator. Otherwise, we think the value of $\|\nabla_{\boldsymbol{z}} G(\boldsymbol{z})\|_p$ is sufficient to bring in an appropriate $\boldsymbol{\delta}$ for the generator. This hinge loss is advantageous over classical squared loss that expects the gradient magnitude to be exactly k, as it is unreasonable to set the same gradient magnitude for data points from distribution $\mathbb{P}_{\boldsymbol{z}}$.

Smooth GAN

So far, we mainly focus on the smoothness of generated images while neglecting their quality. Considering the generation network and the discriminant network within the framework of GANs, we suggest the proposed smooth generator is beneficial for improving the quality of generated images.

Well-trained deep neural networks have been recently found vulnerable to adversarial examples that are imperceptible to human. Most of the studies on adversarial examples are for image classification problem. But in image generation task, we can easily discover failure generations of well-trained generators as well. The noises resulting in these failure cases can thus be regarded as adversarial noise input. WGAN-GP (Arjovsky, Chintala, and Bottou 2017) is a recent promising variant of vanilla GAN,

$$\min_D \mathop{\mathbb{E}}_{\boldsymbol{z} \sim \mathbb{P}_z} [D(G(\boldsymbol{z}))] - \mathop{\mathbb{E}}_{\boldsymbol{x} \sim \mathbb{P}_d} [D(\boldsymbol{x})]. \tag{15}$$

Loss function of WGAN-GP reflects the image quality, which is distinct from loss of vanilla GAN to measure how well it fools the discriminator. The first term in Eq. (15) is relevant to the real sample and has no connection with the generator. Larger value of $D(G(\boldsymbol{z}))$ in Eq. (15) therefore indicates high quality of generated images. If noise vector $\boldsymbol{z}$ generates a high-quality image, we expect that its neighboring point $\boldsymbol{z} + \boldsymbol{\delta}$ would generate an image of high quality as well. To decrease the quality gap between images generated from closed noise inputs, we need to ensure that $D(G(\boldsymbol{z}))$ would not drop significantly when the input variates, *i.e.*,

$$\big|D[G(\boldsymbol{z}+\boldsymbol{\delta})] - D[G(\boldsymbol{z})]\big| < \epsilon. \tag{16}$$

In the following theorem, we analyze what conditions the perturbation $\boldsymbol{\delta}$ should satisfy to guarantee the image quality.

Theorem 2. *Fix $\epsilon > 0$. Consider generator G and discriminator D in GANs. Given a noise input $\boldsymbol{z} \in \mathbb{R}^d$, the generated image is $\hat{\boldsymbol{x}} = G(\hat{\boldsymbol{z}})$. If the perturbation $\boldsymbol{\delta} \in \mathbb{R}^d$ satisfies*

$$\|\boldsymbol{\delta}\|_q < \frac{\epsilon}{\max_{\hat{\boldsymbol{z}}\in B_p(\boldsymbol{z},R)} \|\nabla_{\hat{\boldsymbol{x}}} D(\hat{\boldsymbol{x}})\|_p \|\nabla_{\hat{\boldsymbol{z}}} G(\hat{\boldsymbol{z}})\|_p}, \tag{17}$$

with $\frac{1}{p}+\frac{1}{q}=1$, we have $\big|D[G(\boldsymbol{z}+\boldsymbol{\delta})] - D[G(\boldsymbol{z})]\big| < \epsilon$.

Proof. Without loss of generality, we first suppose $D[G(\boldsymbol{z}+\boldsymbol{\delta})] \geq D[G(\boldsymbol{z})]$. Following the proof of Theorem 1, we can draw a similar conclusion,

$$\begin{aligned} 0 &\leq D[G(\boldsymbol{z}+\boldsymbol{\delta})] - D[G(\boldsymbol{z})] \\ &\leq \|\boldsymbol{\delta}\|_q \max_{\hat{\boldsymbol{z}}\in B_p(\boldsymbol{z},R)} \|\nabla_{\hat{\boldsymbol{z}}} D[G(\hat{\boldsymbol{z}})]\|_p. \end{aligned} \tag{18}$$

According to the chain rule, we have

$$\nabla_{\hat{\boldsymbol{z}}} D[G(\hat{\boldsymbol{z}})] = \nabla_{\hat{\boldsymbol{x}}} D(\hat{\boldsymbol{x}}) \nabla_{\hat{\boldsymbol{z}}} G(\hat{\boldsymbol{z}}), \tag{19}$$

where $\hat{\boldsymbol{x}} = G(\hat{\boldsymbol{z}})$ is the generated image. Given the fact that

$$\|\nabla_{\hat{\boldsymbol{x}}} D(\hat{\boldsymbol{x}}) \nabla_{\hat{\boldsymbol{z}}} G(\hat{\boldsymbol{z}})\|_p \leq \|\nabla_{\hat{\boldsymbol{x}}} D(\hat{\boldsymbol{x}})\|_p \|\nabla_{\hat{\boldsymbol{z}}} G(\hat{\boldsymbol{z}})\|_p, \tag{20}$$

where $\frac{1}{p}+\frac{1}{q}=1$, Eq. (18) can be re-written as

$$\begin{aligned} &\|\boldsymbol{\delta}\|_q \max_{\hat{\boldsymbol{z}}\in B_p(\boldsymbol{z},R)} \|\nabla_{\hat{\boldsymbol{z}}} D[G(\hat{\boldsymbol{z}})]\|_p \\ &\leq \|\boldsymbol{\delta}\|_q \max_{\hat{\boldsymbol{z}}\in B_p(\boldsymbol{z},R)} \|\nabla_{\hat{\boldsymbol{x}}} D(\hat{\boldsymbol{x}})\|_p \|\nabla_{\hat{\boldsymbol{z}}} G(\hat{\boldsymbol{z}})\|_p. \end{aligned} \tag{21}$$

If the right side of Eq. (21) is always upper bounded by ϵ, *i.e.*

$$\|\boldsymbol{\delta}\|_q \max_{\hat{\boldsymbol{z}}\in B_p(\boldsymbol{z},R)} \|\nabla_{\hat{\boldsymbol{x}}} D(\hat{\boldsymbol{x}})\|_p \|\nabla_{\hat{\boldsymbol{z}}} G(\hat{\boldsymbol{z}})\|_p < \epsilon, \tag{22}$$

we then have

$$\|\boldsymbol{\delta}\|_q < \frac{\epsilon}{\max_{\hat{\boldsymbol{z}}\in B_p(\boldsymbol{z},R)} \|\nabla_{\hat{\boldsymbol{x}}} D(\hat{\boldsymbol{x}})\|_p \|\nabla_{\hat{\boldsymbol{z}}} G(\hat{\boldsymbol{z}})\|_p}. \tag{23}$$

In the similar approach, we can get the same constraint (*i.e.*, Eq. (23)) over $\boldsymbol{\delta}$ to achieve $-\epsilon < \big|D[G(\boldsymbol{z}+\boldsymbol{\delta})] - D[G(\boldsymbol{z})]\big|$. The proof is completed. □

Based on Theorem 2, we propose to minimize

$$\max_{\hat{\boldsymbol{z}}\in B_p(\boldsymbol{z},R)} \|\nabla_{\hat{\boldsymbol{x}}} D(\hat{\boldsymbol{x}})\|_p \|\nabla_{\hat{\boldsymbol{z}}} G(\hat{\boldsymbol{z}})\|_p, \tag{24}$$

so that the upper bound over $\boldsymbol{\delta}$ will be enlarged and GANs model is expect to tolerate more drastic perturbation. Since it is difficult to discover the optimal $\hat{\boldsymbol{z}} \in B_p(\boldsymbol{z}, R)$, we suppose that there is an approximated $\hat{\boldsymbol{z}}$ sampled from the distribution $\mathbb{P}_z$ as well. The loss function can then be reformulated as,

$$\min_{G,D} \mathop{\mathbb{E}}_{\boldsymbol{z}\sim\mathbb{P}_z} \|\nabla_{\overline{\boldsymbol{x}}} D(\overline{\boldsymbol{x}})\|_p \|\nabla_{\boldsymbol{z}} G(\boldsymbol{z})\|_p, \tag{25}$$

where $\overline{\boldsymbol{x}} \sim \mathbb{P}_G$ is the generated sample $G(\boldsymbol{z})$. Two terms $\|\nabla_{\boldsymbol{x}} D(\boldsymbol{x})\|_p$ and $\|\nabla_{\boldsymbol{z}} G(\boldsymbol{z})\|_p$ are involved in Eq. (25). WGAN-GP (Gulrajani et al. 2017) proposed gradient-penalty,

$$\mathcal{L}_{GPoD} = \mathop{\mathbb{E}}_{\tilde{\boldsymbol{x}}\sim\mathbb{P}_x} [(\|\nabla_{\tilde{\boldsymbol{x}}} D(\tilde{\boldsymbol{x}})\|_2 - 1)^2], \tag{26}$$

where $\mathbb{P}_x$ consists of both real sample distribution $\mathbb{P}_d$ and generated sample distribution $\mathbb{P}_G$. By concentrating on generated samples $\overline{\boldsymbol{x}} \sim \mathbb{P}_G$, Eq. (26) encourages $\|\nabla_{\overline{\boldsymbol{x}}} D(\overline{\boldsymbol{x}})\|_2$ to go towards 1, and has been proved to successfully constrain the norm of the gradient of discriminator $\|\nabla_{\overline{\boldsymbol{x}}} D(\overline{\boldsymbol{x}})\|_2$ in experiments. The remaining term $\|\nabla_{\boldsymbol{z}} G(\boldsymbol{z})\|_p$ in Eq. (25) is therefore our only focus. In a similar approach, we encourage the norm of the gradient of generator to stay at a lower level and reformulate Eq. (25) to

$$\mathcal{L}_{GPoG} = \mathop{\mathbb{E}}_{\boldsymbol{z}\sim\mathbb{P}_z} \max(0, \|\nabla_{\boldsymbol{z}} G(\boldsymbol{z})\|_2^2 - k), \tag{27}$$

where we set $p = q = 2$. This equation is exactly the same as Eq. (14).

Algorithm 1 Smooth GAN

```
Require: The number of critic iterations per generator iteration $n_{critic}$, the batch size $m$, Adam hyperparameters $\alpha$, $\beta_1$, and $\beta_2$, the loss balanced coefficient $\lambda$, $\gamma$.
Require: initial discriminator parameters $w_0$, initial generator parameters $\theta_0$.
repeat
for $t = 1, ..., n_{critic}$ do
    for $i = 1, ..., m$ do
        Sample real data $\boldsymbol{x} \sim \mathbb{P}_d$, latent variable $\boldsymbol{z} \sim \mathbb{P}_z$, a random number $t \sim U[0, 1]$.
        Calculate fake sample $G_\theta(\boldsymbol{z})$, interpolation sample $\tilde{\boldsymbol{x}} \leftarrow t\boldsymbol{x} + (1-t)G_\theta(\boldsymbol{z})$
        Calculate the loss function $\mathcal{L}_D^{(i)} \leftarrow D_w[G_\theta(\boldsymbol{z})] - D_w(\boldsymbol{x}) + \lambda(\|\nabla_{\tilde{\boldsymbol{x}}} D(\tilde{\boldsymbol{x}})\|_2 - 1)^2$;
    end for
    Update discriminator parameters
    $w \leftarrow Adam(\nabla_w \frac{1}{m}\sum_{i=1}^m L_D^{(i)}, w, \alpha, \beta_1, \beta_2)$
end for
Sample a batch of latent variables $\{\boldsymbol{z}^{(i)}\}_{i=1}^m \sim \mathbb{P}_z$.
Calculate the loss function
$L_G^{(i)} \leftarrow -D_w[G_\theta(\boldsymbol{z})] + \gamma \max(0, \|\nabla_{\boldsymbol{z}} G(\boldsymbol{z})\|_2^2 - k)$;
Update generator parameters
$\theta \leftarrow Adam(\nabla_\theta \frac{1}{m}\sum_{i=1}^m L_G^{(i)}, \theta, \alpha, \beta_1, \beta_2)$
until $\theta$ has converged
Ensure: A smooth generator network $G$.
```

By integrating Eqs. (26) and (27) with WGAN, we obtain the resulting objective function:

$$\begin{aligned} \mathcal{L} = &\mathop{\mathbb{E}}_{\boldsymbol{z}\sim\mathbb{P}_z}[D(G(\boldsymbol{z}))] - \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathbb{P}_d}[D(\boldsymbol{x})] \\ &+ \lambda \mathop{\mathbb{E}}_{\tilde{\boldsymbol{x}}\sim\mathbb{P}_x}[(\|\nabla_{\tilde{\boldsymbol{x}}} D(\tilde{\boldsymbol{x}})\|_2 - 1)^2] \\ &+ \gamma \mathop{\mathbb{E}}_{\boldsymbol{z}\sim\mathbb{P}_z} \max(0, \|\nabla_{\boldsymbol{z}} G(\boldsymbol{z})\|_2^2 - k), \end{aligned} \tag{28}$$

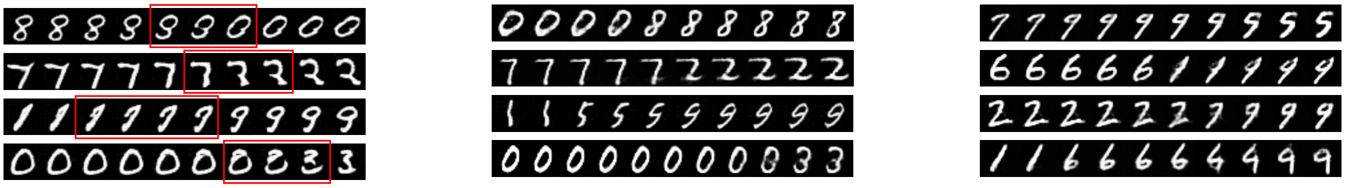


Figure 2: Illustration of image interpolations on the MNIST dataset generated from WGAN-GP (Gulrajani et al. 2017) (left) and the proposed method (middle and right).

where λ and γ constant numbers to balance different terms in the function. Our complete algorithm pipeline is summarized in Algorithm 1.

Experimental Results

In this section, we conduct comprehensive experiments on a toy dataset and three real-world image datasets, MNIST (LeCun et al. 1998), CIFAR-10 (Krizhevsky and Hinton 2009), and CelebA (Liu et al. 2015).

Datasets and Settings

In this part, we introduce the real image datasets used in the experiments and the corresponding experimental settings and network structure. In addition, all the images are normalized to pixel values in $[-1, +1]$. We utilize different network architectures on different datasets that was detailed in following. The common points are: i) no nonlinear activation was attached to the end of discriminators; ii) the minibatch used in training process is 64 for both generator and discriminators; iii) Adam optimizer with learning rate 0.0001 and momentum 0.5; iv) noise dimension of 128 for generator; v) weights initialized from Gaussian: $\mathcal{N}(0; 0.01)$.

MNIST (LeCun et al. 1998) is a handwritten digits dataset (from 0 to 9) composed of 28×28 pixel greyscale images from ten categories. The whole dataset of 70,000 images is split into 60,000 and 10,000 images for training and test, respectively. In the experiments on the MNIST dataset, we consider the 10,000 images in test set as valid set in the calculation of FID.

CIFAR-10 (Krizhevsky and Hinton 2009) is a dataset that consists of 32×32 pixel RGB color images drawn from 10 categories. There are 60,000 images in the CIFAR-10 dataset which are split into 50,000 training and 10,000 testing images. We also calculate the FID with 3,000 images that was randomly selected in the test set.

CelebA (Liu et al. 2015) is a dataset consist of 202,599 portraits of celebrities. We use the aligned and cropped version, which preprocesses each image to a size of 64×64 pixels. 3,000 examples are randomly selected as the test set and the rest samples as the training set.

Evaluation Metrics

We evaluate the proposed method mainly in terms of n three metrics well suited to the image domain.

Inception score (IS) (Salimans et al. 2016) rewarding high-quality and high-variability of samples, can be expressed as: $\exp(\mathbb{E}_{\boldsymbol{x}}[D_{KL}(p(y|\boldsymbol{x})\|p(y))])$, where $p(y) = \frac{1}{N}\sum_{i=1}^{N} p(y|\boldsymbol{x}^i = G(\boldsymbol{z}^i))$ is the margin distribution and $p(y|\boldsymbol{x})$ is the conditional distribution for samples. In this paper, we estimate the IS using a Inception model (Szegedy et al. 2016) pretrained in torchvision of PyTorch.

Frechet Inception Distance (FID) (Heusel et al. 2017) described the distance between two distributions. FID is computed as follow:

$$\mathrm{FID} = \|\mu_{\mathrm{g}} - \mu_{\mathrm{r}}\|_2^2 + \mathrm{Tr}(\Sigma_{\mathrm{g}} + \Sigma_{\mathrm{r}} - 2(\Sigma_{\mathrm{g}}\Sigma_{\mathrm{r}})^{\frac{1}{2}}), \quad (29)$$

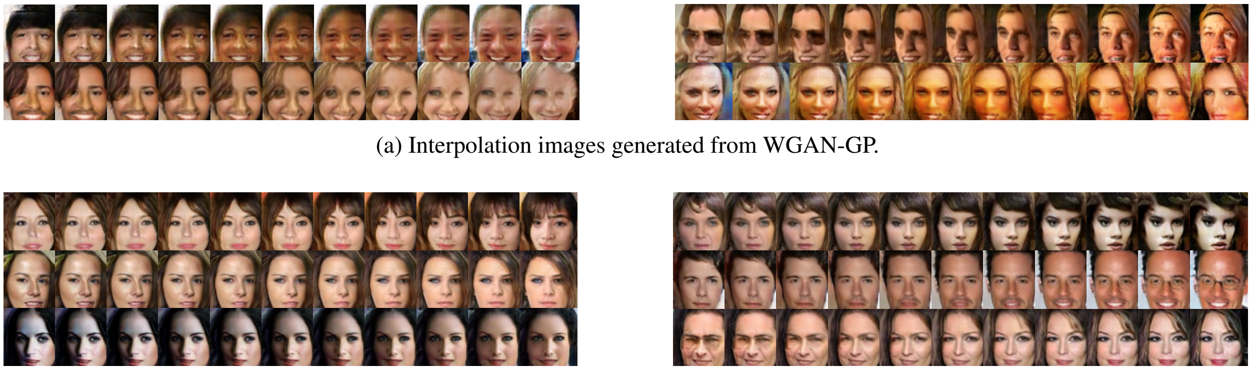
where (μ_g, Σ_g) and (μ_r, Σ_r) are the mean and covariance of embedded samples from generated distribution $\mathbb{P}_g$ and real image distribution $\mathbb{P}_r$, respectively. In the paper, we regard the feature maps obtained from a specific layer of the pretrained Inception model as the embedding of the samples. FID is more sensitive to the diversity of samples belonging to the same category and fixes the drawback of inception score that is easily fooled by a model which generated only one image per class. We describe the quality of models together with the FID and IS score.

Multi-scale Structural Similarity for Image Quality (MS-SSIM) (Wang, Simoncelli, and Bovik 2003) was proposed to measuring the similarity of two images. This metric well suits to evaluate the quality of samples belonging to one class. As a result, we apply it to describe the smoothness of samples obtained from the interpolation opretion. If two algorithms receive similar FID and IS socores, the samples could be equivalent on quality and diversity, and higher MS-SSIM score means a smoother conversion process.

Interpolation

We implement the images interpolation on the MNIST and CelebA datasets and show a few results in Figures (2) and (3). We forward a series of noises obtained by linearly interpolating between two points in noise space into the model and expect that the resulting images show smooth transposition in the image space.

In Figure 2, the left part and the middle part showing the similar transposition are generated from models trained with WGAN-GP and the Smooth GAN, respectively. The left part shows more meaningless images (high lighted in red) during the transposition. When changing an image from one class to the other one, we want to keep the whole process smooth and meaningful. As shown in the right part of Figure 2, these images accomplish these transpositions by considering image qualities and image semantics. For example, the number '6' becomes '1' firstly and then becomes '9'. Obviously, '1' is more closed to '9' than '6'. Interpolation results on the



(a) Interpolation images generated from WGAN-GP.

(b) Interpolation images generated from the proposed Smooth GAN.

Figure 3: Image interpolations on the CelebA dataset.

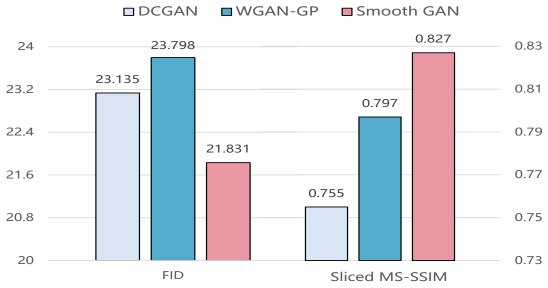


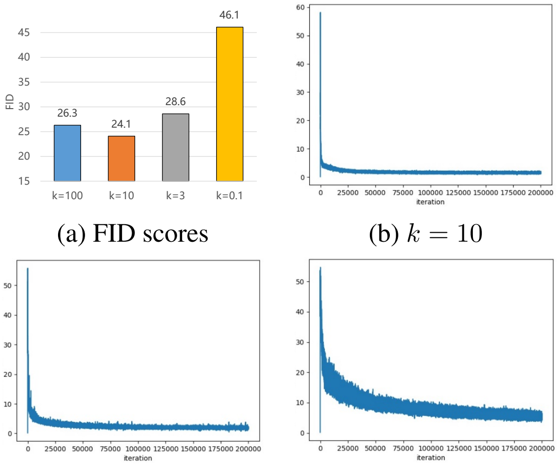
Figure 4: FID and sliced MS-SSIM obtained by different models on the CIFAR-10 dataset.

other numbers also illustrate this phenomenon. In the results on the CelebA dataset shown in Figure 3, we achieve great quality while maintaining smoothness, which illustrates the effectiveness of our approach.

Taken two images and their interpolation images generated from the linear interpolations in noise space as a interpolation slice, to illustrate the effectiveness of the proposed method in a more convincing way, we generate several slides on the CelebA dataset. We generated such slices based on different models and calculated the FID and MS-SSIM of these slices for comparison. Different from calculating the MS-SSIM score over whole resulting images, we calculated it for every interpolation slice independently. Higher MS-SSIM values correspond to perceptually more similar images but also lower diversity and mode collapse (Odena, Olah, and Shlens 2016; Fedus et al. 2017). Meanwhile, higher FID score ensures the diversity and prevents the GANs from mode collapse. As a result, considering together with FID, the MS-SSIM score could focus on indicating the similarity of images, which is consistent with smooth transposition in an interpolation slide. The sliced MS-SSIM score used in this experiment can be described as:

$$\frac{N(N-1)}{2}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\mathrm{MS}-\mathrm{SSIM}(S_i,S_j), \tag{30}$$

where S_i is the i-th slide of samples in the resulting group.



(a) FID scores

(b) $k = 10$

(c) $k = 3$

(d) $k = 0.1$

Figure 5: FID scores (a) and Wasserstein distance convergence (b, c, d) under different values of k.

Now, we could estimate the effectiveness with FID and sliced MS-SSIM. We report the FID and sliced MS-SSIM obtained on the CelebA dataset in Figure 4. Our model not only has the lowest FID score, but also its MS-SSIM score exceeds all other models. This is consistent with our observations and demonstrates the effectiveness of our approach.

Hyperparemeter Analysis To illustrate the choice of the value of k, we show FID scores and Wasserstein distance convergence curves of experiments with different k values on the CIFAR-10 dataset in Figure 5. Figure 5 (a) shows the FID scores obtained from four k values, and $K = 10$ provides the best score. Figure 5 (b) shows that the experiment with $k = 10$ achieves the best convergence. Setting k to 3 will influence the training progress and achieve slightly higher Wasserstein distance than Figure 5 (b) when network convergence. The generator is failed to produce enough realistic images when setting k to 0.1; this is because too small a k value will suppress the diversity of generator output.

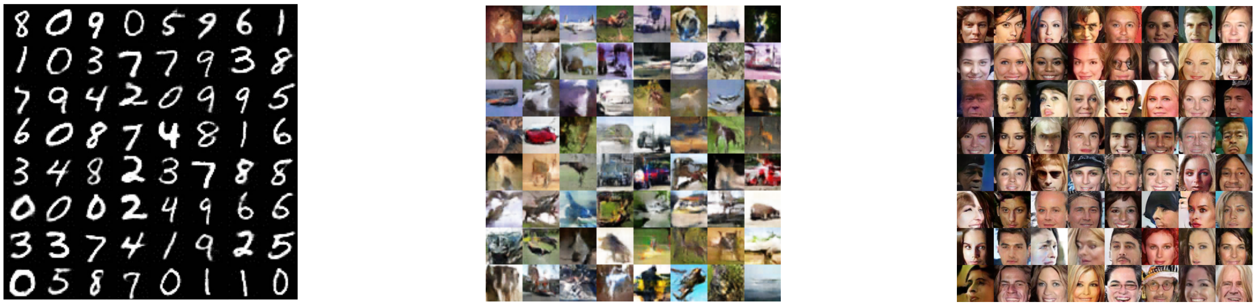


Figure 6: Samples generated from the proposed model trained on MNIST (left), CIFAR-10 (middle), and CelebA (right).

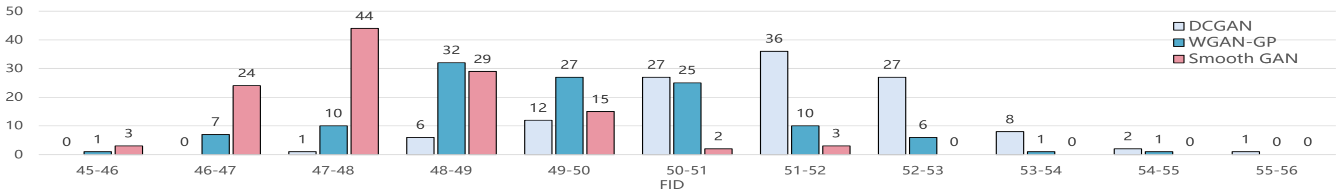


Figure 7: FID distribution obtained on several subsets.

Moreover, a large value of k relaxes the constraint on the gradient and can result in insufficient smoothness of the network output. In our experiments, we used different k values based on experimental results regarding different datasets.

Table 1: Incetion scores (higher is better) and FIDs (lower is better) on the CIFAR-10 dataset.

Method	IS ↑	FID ↓
DCGAN (Radford, Metz, and Chintala 2015)	$6.40 \pm .05$	36.7
BEGAN (Berthelot, Schumm, and Metz 2017b)	$5.62 \pm .07$	32.3
WGAN (Arjovsky, Chintala, and Bottou 2017)	$3.82 \pm .06$	42.8
D2GAN (Nguyen et al. 2017)	$7.15 \pm .07$	30.6
WGAN-GP (Gulrajani et al. 2017)	$7.36 \pm .07$	26.5
Smooth GAN (Ours)	$\mathbf{7.66 \pm .05}$	**24.1**

Image Generation

We conduct experiments on three real image datasets to investigate the capabilities of the proposed method. Table 1 reports the inception scores and FIDs on the CIFAR-10 dataset which obtained from the proposed model and baseline methods. In this results, the proposed method outperforms almost state-of-the-art methods. Therefore, the proposed method provides considerable quality on the three datasets.

Figure 6 shows several samples generated by the model learned with the proposed method. The samples on the MNIST dataset show a variety of numbers and styles. Dogs, trucks, boats, and fish could also be found in the samples on the CIFAR-10 dataset. Age and gender diversity can also be observed in the results on the CelebA dataset. These results confirm the capabilities of the proposed method.

Samples Quality Distribution

In this section, we introduce a way to describe the quality distribution of samples and demonstrate the effectiveness of the proposed method that can effectively reduce the number of low-quality images. FID is a good indicator to evaluate the quality of the generated samples. However, FID only provides an average quality of the whole test images. First, we generate a sufficient large set of images (50,000 in experiments). Next, we randomly sample 512 images to form a subset and calculate the FID of this subset. We repeat the second step 120 times and calculate their FID scores. By comparing these FIDs obtained from subsets, we can roughly estimate the quality distribution of the samples. Figure 7 shows the distribution of FID scores calculated over subsets and obtained from three models. Compared to other models, our model can obtain lower FID scores, while bigger value of FIDs are less. Moreover, the FID scores obtained from the proposed model are mainly concentrated between 46 and 50, and only 8 scores of subset fall outside this region, while the other two algorithms get a loose distribution of FID scores. Therefore, our method can effectively reduce the low-quality samples in the generated samples.

Conclusions

Here we analyze the relationship between perturbation on the input of the generator and its influence on the output images. By investigating the tolerable input perturbation, we develop a smooth generator. We further integrate this smooth generator with a gradient penalized discriminator, and produce smooth GAN that generate stable and high-quality images. Experiments on real-world image datasets demonstrate the necessity of studying smooth generator and show the proposed method is capable of learning smooth GAN.

Acknowledgments

We thank supports of the NSFC under Grant 61876007 and 61872012, ARC Projects FL-170100117, DP-180103424, IH-180100002, DE-180101438, and Beijing Major Science and Technology Project Z171100000117008.

References

Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein gan. *arXiv preprint arXiv:1701.07875*.

Berthelot, D.; Schumm, T.; and Metz, L. 2017a. Began: boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717*.

Berthelot, D.; Schumm, T.; and Metz, L. 2017b. BEGAN: boundary equilibrium generative adversarial networks. *CoRR* abs/1703.10717.

Bojanowski, P.; Joulin, A.; Lopez-Paz, D.; and Szlam, A. 2017. Optimizing the latent space of generative networks. *arXiv preprint arXiv:1707.05776*.

Chen, X.; Duan, Y.; Houthooft, R.; Schulman, J.; Sutskever, I.; and Abbeel, P. 2016. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems*, 2172–2180.

Chen, X.; Xu, C.; Yang, X.; and Tao, D. 2018. Attention-gan for object transfiguration in wild images. In *The European Conference on Computer Vision (ECCV)*.

Chen, X.; Xu, C.; Yang, X.; Song, L.; and Tao, D. 2019. Gated-gan: Adversarial gated networks for multi-collection style transfer. *IEEE Transactions on Image Processing* 28(2):546–560.

Denton, E. L.; Chintala, S.; Szlam, A.; and Fergus, R. 2015. Deep generative image models using a laplacian pyramid of adversarial networks. *CoRR* abs/1506.05751.

Fedus, W.; Rosca, M.; Lakshminarayanan, B.; Dai, A. M.; Mohamed, S.; and Goodfellow, I. 2017. Many paths to equilibrium: Gans do not need to decrease adivergence at every step. *arXiv preprint arXiv:1710.08446*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, 2672–2680.

Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; and Courville, A. C. 2017. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, 5767–5777.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; Klambauer, G.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a nash equilibrium. *arXiv preprint arXiv:1706.08500*.

Karras, T.; Aila, T.; Laine, S.; and Lehtinen, J. 2017. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*.

Kim, T., and Bengio, Y. 2016. Deep directed generative models with energy-based probability estimation. *arXiv preprint arXiv:1606.03439*.

Kingma, D. P., and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Krizhevsky, A., and Hinton, G. 2009. Learning multiple layers of features from tiny images.

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* 86(11):2278–2324.

Li, C.; Xu, K.; Zhu, J.; and Zhang, B. 2017a. Triple generative adversarial nets. *arXiv preprint arXiv:1703.02291*.

Li, C.-L.; Chang, W.-C.; Cheng, Y.; Yang, Y.; and Póczos, B. 2017b. Mmd gan: Towards deeper understanding of moment matching network. In *Advances in Neural Information Processing Systems*, 2203–2213.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, 3730–3738.

Makhzani, A.; Shlens, J.; Jaitly, N.; Goodfellow, I.; and Frey, B. 2015. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*.

Mao, X.; Li, Q.; Xie, H.; Lau, R. Y.; Wang, Z.; and Smolley, S. P. 2017. Least squares generative adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2813–2821. IEEE.

Mirza, M., and Osindero, S. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.

Nguyen, T.; Le, T.; Vu, H.; and Phung, D. 2017. Dual discriminator generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2670–2680.

Odena, A.; Olah, C.; and Shlens, J. 2016. Conditional image synthesis with auxiliary classifier gans. *arXiv preprint arXiv:1610.09585*.

Radford, A.; Metz, L.; and Chintala, S. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; and Chen, X. 2016. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, 2234–2242.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826.

van den Oord, A.; Kalchbrenner, N.; Espeholt, L.; Vinyals, O.; Graves, A.; et al. 2016. Conditional image generation with pixelcnn decoders. In *Advances in Neural Information Processing Systems*, 4790–4798.

Wang, Z.; Simoncelli, E. P.; and Bovik, A. C. 2003. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, 1398–1402. Ieee.

Zhai, S.; Cheng, Y.; Feris, R.; and Zhang, Z. 2016. Generative adversarial networks as variational training of energy based models. *arXiv preprint arXiv:1611.01799*.