Indirect Adversarial Losses via an Intermediate Distribution for Training GANs

Rui Yang Duc Minh Vo Hideki Nakayama
{yang, vmduc}@nlab.ci.i.u-tokyo.ac.jp nakayama@ci.i.u-tokyo.ac.jp
The University of Tokyo, Japan

(a) Traditional adversarial loss (b) Indirect adversarial loss

Figure 1: Illustration of traditional or our indirect adversarial loss. 'D step' and 'G step' refer to the discriminator and the generator steps, respectively. In past work, the adversarial learning process was directly connected between the discriminator outputs of real and fake samples, while we construct an indirect process among real, fake, and intermediate distribution to avoid the attractive problem.

Abstract

In this study, we consider the weak convergence characteristics of the Integral Probability Metrics (IPM) methods in training Generative Adversarial Networks (GANs). We first concentrate on a successful IPM-based GAN method that employs a repulsive version of the Maximum Mean Discrepancy (MMD) as the discriminator loss (called repulsive MMD-GAN). We reinterpret its repulsive metrics as an indirect discriminator loss function toward an intermediate distribution. This allows us to propose a novel generator loss via such an intermediate distribution based on our reinterpretation. Our indirect adversarial losses use a simple known distribution (i.e., the Normal or Uniform distribution in our experiments) to simulate indirect adversarial learning between three parts – real, fake, and intermediate distributions. Furthermore, we found the Kernelized Stein Discrepancy (KSD) from the IPM family as the adversarial loss function to avoid randomness from intermediate distribution samples because the target side (intermediate one) is sample-free in KSD. Experiments on several real-world datasets show that our methods can successfully train GANs with the intermediate-distribution-based KSD and MMD and can outperform previous loss metrics.

1. Introduction

Although the Generative Adversarial Networks (GANs) [7] have been highly successful, training GANs remains challenging. To tackle this problem, multiple strategies have been proposed such as designing loss functions [1, 20], network archi-

tectures [12, 14], and training regularization [26, 9]. Compared with f-divergence families [3, 21, 30], Integral Probability Metrics (IPM) [29] GANs [1, 9, 20, 35] (IPM-GANs) imply weak convergence, achieving higher generation quality [24]. Theoretically, IPM methods can reach the numerical zero between fake and real distributions if and only if two distributions are equal [24] (Here, the real and fake *distributions* refer to the posterior outputs of the discriminator conditioned on the discriminator inputs, later the same). For instance, the Maximum Mean Discrepancy (MMD) shows excellent performance in MMD-GAN [20]. Latter, the repulsive MMD-GAN [35] changed the discriminator loss from an attractive MMD to a repulsive MMD discriminator loss and using the same generator loss as MMD-GAN.

In this paper, we attempt to solve the attractive problem as shown in in repulsive MMD-GAN [35] without the mixed use of two different loss metrics (attractive and repulsive losses), whose learning directions are not unified (G steps as Fig. 1^a while D steps as Fig. 1^b). Furthermore, we derive a new ideal framework for solving such a mixing problem. We reinterpret the repulsive MMD discriminator loss in Eq. 3 without including the repulsiveness. Precisely, we rewrite the repulsive MMD equation via a pseudo intermediate distribution as the learning target for the output distribution. The attractive MMD generator loss in the original MMD-GAN [20] directly moves the fake distribution towards the real distribution. In our explanation of the repulsive MMD discriminator loss [35], the real distribution is moved towards the pseudo intermediate distribution, which cannot unify the min-max game. Thus, we propose a novel generator loss to pair with the repulsive MMD discriminator loss that can avoid the mixing problem via our indirect MMD losses.

We maintain a dynamic balance between fake and real distributions near a known intermediate distribution. This learning process realizes indirect adversarial learning among three distributions – fake, real, and intermediate. Specifically, the real distribution moves close to the intermediate distribution while the fake one moves far from the intermediate one in the discriminator steps. Then, the fake distribution moves toward the intermediate distribution in the generator steps. After training, the fake and real distributions are close to the intermediate one, thereby minimizing the real and fake distributions.

Moreover, computing the MMD distance requires random samples from the intermediate distribution, resulting in sample-based bias during the training process. Thus, we exploit a specific sample-free IPM method, namely, Stein Discrepancy (SD) [36], to avoid such randomness. SD and its kernelized version (KSD) [22, 4, 5] have been widely applied to many machine learning tasks, such as variational auto-encoder [32], artificial sampler [11], and energy-based models [8], but not yet to training a data-driven GANs. We propose to replace the loss functions used in previous work with our novel KSD-based loss functions to solve the randomness problem in our indirect version of MMD-GAN.

Our simple yet effective method combines the key idea of adversarial learning and KSD, namely KSD-GAN, notably improving the training of GANs in terms of generation quality. Our contributions are as follow:

- We propose an indirect adversarial training process to unify the generator and discriminator losses in repulsive MMD-GAN (Fig. 1).
- We found KSD losses to overcome the randomness in the indirect version of repulsive MMD losses to improve the learning process.
- Our real-world datasets experiments showed superior performances to other loss metrics.

2. Background

2.1. IPM GANs

IPMs [29] were defined to maximize the difference between the expectations of the source distribution p and the target distribution q via the witness function $f(\cdot)$:

$$IPM_{\mathcal{F}}(p,q) = \sup_{f\in\mathcal{F}} \left| \int f\, dp - \int f\, dq \right|, \quad (1)$$

where $\mathcal{F}$ in Eq. 1 is a class of real-valued bounded measurable functions.

Depending on the different conditions of the witness function $f(\cdot)$, the IPM family involves many types of measurements. For instance, the Wasserstein-1 distance requires the Lipschitz continuity of function $f(\cdot)$. Another typical example is MMD, which defines the witness function in the RKHS. Besides Wasserstein and MMD, other definitions also matter, such as the Fisher IPM, and SD, etc. As a result, significant successes have been achieved with GANs based on the above mentioned IPM definitions as the adversarial losses [1, 20, 27].

2.2. Maximum Mean Discrepancy and Repulsive Loss

The squared MMD defines the difference between the source distribution p and the target distribution q based on the kernel function $k(\cdot)$. As shown in [28], a mathematical approximation can be used to compute the MMD distance numerically:

$$\begin{aligned} MMD(p,q) = &\mathbb{E}_{y\sim q}[k(D(y),D(y'))] \\ &- 2 * \mathbb{E}_{x\sim p, y\sim q}[k(D(x),D(y))] \\ &+ \mathbb{E}_{x\sim p}[k(D(x),D(x'))], \end{aligned} \tag{2}$$

where $D(\cdot)$ stands for the discriminator outputs, and x and y are samples from source distribution p and the target distribution q, respectively.

In MMD-GAN [20], Eq. 2 is used as the generator loss while the negative form of Eq. 2 is used as the discriminator loss. Subsequently, the authors who proposed repulsive MMD-GAN [35] showed that negatively using Eq. 2 as the discriminator loss leads to a smaller intra-distance among discriminator outputs in real samples, which is called the attractive problem. Thus, they proposed the repulsive version of the MMD function as:

$$\begin{aligned} MMD_{rep}(p,q) = &\mathbb{E}_{x\sim p}[k(D(x),D(x'))] \\ &- \mathbb{E}_{y\sim q}[k(D(y),D(y'))]. \end{aligned} \tag{3}$$

The repulsive MMD-GAN also uses Eq. 2 as the generator loss to pair with Eq. 3. In [7], the convergence of mini-max game was proved based on JS-divergence. Replacing JS-divergence with the MMD distance does not change the conclusion in [7], while different loss metrics in the generator and discriminator steps were used in [35]. Thus, the conclusion of the convergence from vanilla-GAN cannot be used directly in repulsive MMD-GAN.

2.3. Stein Discrepancy

Stein Discrepancy is derived from the goodness-of-fitting test [36], which is a special case in IPM methods. The definition of the score function in SD infers Stein's Identity [22]:

$$\mathbb{E}_p\left[S_q(x)f(x)^\top + \bigtriangledown_x f(x)\right] = 0, \tag{4}$$

where the score function is $S_q(x) = \bigtriangledown_x log(q(x))$ and $q(x)$ is the unnormalized p.d.f. of the target distribution. The smoothness requirement of each f in Eq. 4 is the same as that in Eq. 1 and all $f \in \mathcal{F}$ are in the Stein class of the p distribution. Eq. 4 can be satisfied if and only if $p = q$.

Thus, Stein's method can indicate how well a given set of samples matches a specific target distribution. The measure $\mathbb{S}(p,q)$ between samples from the source distribution and p.d.f. in the score function S_q is defined as:

$$\sup_{f\in\mathcal{F}}\left\{\mathbb{E}_{x\sim p}[S_q(x)f(x)^\top + \bigtriangledown_x f(x)]\right\}, \tag{5}$$

where p is the source distribution and q is the target distribution. It is clear that $\mathbb{S}(p,q)$ only depends on the samples in the source distribution.

By introducing SD into RKHS, the authors in [22] solved the computational problem, making it feasible for becoming a statistical loss function in many machine learning tasks. As shown in [22], KSD can be written in a kernelized form, $\mathbb{E}_{x,x'\sim p}[u_q(x,x')]$, where the u_q kernel can be extended as:

$$\begin{aligned} u_q(x,x') = &S_q(x)^\top k(x,x')S_q(x') \\ &+ S_q(x)^\top \bigtriangledown_x k(x,x') \\ &+ \bigtriangledown_x k(x,x')^\top S_q(x') \\ &+ tr(\bigtriangledown_{x,x'} k(x,x')). \end{aligned} \tag{6}$$

Here, the choice of the $k(x,x')$ kernel was the RBF kernel in all kernel-related experiments owing to its empirical performances in past works [22, 23].

3. Indirect Adversarial Losses

3.1. Reinterpreting Repulsive MMD loss

In [35], the discriminator removed the intersection kernel matrix (the second term in Eq. 2) between the real and fake discriminator outputs distributions. First, we define a pseudo target distribution (denoted as $\mathcal{O}$), that only contains samples equal to zero, namely, the Dirac-δ distribution as in Fig. 1. Then, we can replace the repulsive MMD discriminator loss function in terms of this intermediate distribution $\mathcal{O}$: $L_{\mathcal{D}} = MMD(p, \mathcal{O}) - MMD(q, \mathcal{O})$, which moves p close to the intermediate distribution $\mathcal{O}$ and moves the target distribution q away from $\mathcal{O}$ in accordance with MMD distance. Because $\mathcal{O}$ only contains zero, both $MMD(p, \mathcal{O})$ and $MMD(q, \mathcal{O})$ have constant matrixes in the third term in Eq. 2. Constant terms only give zero gradients in backpropagation, allowing us to omit them. The second term in Eq. 2 is the kernel matrix between inputs and zeros; thus, it can be treated as a part of the regularization in the loss function. In this case, because the discriminator loss minimizes the MMD distance between real sample outputs and the intermediate distribution, so no need to negatively use Eq.2, thus no attractive problem happens. Hence, we proposed another repulsive version of the MMD discriminator loss function as an indirect loss function among the real, fake, and intermediate distributions.

$$\begin{aligned} L_{\mathcal{D}} = &\mathbb{E}_{r,r' \sim real}[k(\mathcal{D}(r), \mathcal{D}(r'))] \\ &- \mathbb{E}_{g,g' \sim fake}[k(\mathcal{D}(g), \mathcal{D}(g'))] \\ &- 2 * \mathbb{E}_{r \sim real}[k(\mathcal{D}(r), 0)] \\ &+ 2 * \mathbb{E}_{g \sim fake}[k(0, \mathcal{D}(g))]. \end{aligned} \tag{7}$$

Therefore, we can generalize such indirect loss function to the generator loss based on the intermediate distribution:

$$\begin{aligned} L_{\mathcal{G}} = &\mathbb{E}_{g \sim fake}[k(\mathcal{D}(g), \mathcal{D}(g'))] \\ &- 2 * \mathbb{E}_{g \sim fake}[k(0, \mathcal{D}(g))]. \end{aligned} \tag{8}$$

Our novel generator loss in Eq. 8 has the same learning target as the discriminator loss in repulsive MMD-GAN (as Fig. 1^b). Moreover, compared with chasing the discriminator output distribution from real samples directly in the generator step, targeting $\mathcal{O}$ distribution can suppress circling when the generator falls into a local minima since the optimizer cannot ensure a better network in every step. More importantly, we can replace Dirac-δ distribution with another simple distribution, such as standard Normal distribution ($\mathcal{N}$).

In contrast, our generator loss may face a cold-start problem due to the discriminator having to sample sufficient random samples for the intermediate distribution. As a result, it is hard to acquire meaningful information for the generator during the initial steps. On the other hand, we found another elegant way to solve this issue, called KSD. Owing to its sample-free feature, we can treat the intermediate distribution as the target side q in Eq. 6. Consequently, our proposed KSD-GAN can keep the merits of indirect losses without facing randomness in sampling.

3.2. KSD Loss Function

Firstly, we choose a simple distribution as the intermediate distribution q, i.e., Normal or Uniform distribution. Then, in the discriminator step, we force the discriminator outputs from the real inputs to move close to the intermediate distribution and dissociate the fake ones via KSD to strengthen the discriminating ability. Next, in the generator step, we move the outputs of the discriminator from the fake source close to the intermediate distribution. Finally, the real and fake distributions achieve a dynamic balance near the intermediate distribution. The objective functions are as follows:

$$\begin{aligned} L_{\mathcal{D}} = &\mathbb{E}_{r,r' \sim real}[u_q(\mathcal{D}(r), \mathcal{D}(r'))] \\ &- \mathbb{E}_{g,g' \sim fake}[u_q(\mathcal{D}(g), \mathcal{D}(g'))], \\ L_{\mathcal{G}} = &\mathbb{E}_{g,g' \sim fake}[u_q(\mathcal{D}(g), \mathcal{D}(g'))], \end{aligned} \tag{9}$$

where u_q kernel was defined in Eq. 6 and applied to the outputs of the discriminator $D(\cdot)$. Here, our discriminator loss function minimizes KSD distance between the real outputs and the intermediate distribution while maximizing that between the fake outputs and the intermediate distribution. The KSD distance between the fake outputs and the intermediate distribution is minimized in the generator

loss function. Therefore, the ideal case for the perfectly trained generator will be $KSD(D(r), q) = KSD(D(g), q) = 0$. In this case, the real distribution is equivalent to the generated fake distribution under the measure of KSD. We summarize all related the loss metrics in Table 1.

3.3. Convergence

Our intermediate-distribution-based MMD and KSD-based losses differ from past adversarial losses. We prove their convergence property in two steps: first, our methods belong to IPM-GANs; second, they satisfy the conditions of IPM-GANs regarding convergence, thus obtaining the proof.

Lemma 1. *Define a real-valued bounded measurable function as the witness function $f^*(\cdot)$. Let the q distribution be the intermediate distribution. Then the object function of the intermediate-distribution-based adversarial divergence in IPM-GANs is:*

$$\inf_{g} \sup_{f^* \in \mathcal{F}} \left| \mathbb{E}_{r \sim real, g \sim fake}[f_q^*(r) + f_q^*(g)] \right|. \quad (10)$$

Proof. See Appendix A. □

Corollary 2. *Our intermediate-distribution-based adversarial divergence satisfies the convergence conditions [24] of IPM-GANs. Thus, our methods can obtain convergence.*

3.4. Choice of Intermediate Distribution

The choice of the intermediate distribution for training GANs affects the model performance, with a better balance achieved between quality and diversity by the selection of distributions for datasets on different scales. Normal ($\mathcal{N}$) and Uniform ($\mathcal{U}$) distributions have concise equations and are suitable intermediate distributions. In contrast, some complex distributions, such as mixed Gaussian and Dirichlet distribution, are not suitable for intermediate distributions.

4. Experiments

4.1. Preliminaries

We first tested our intermediate-distribution-based methods on MMD-based and KSD-based losses. We trained CIFAR10 to compare unconditional generation performances, the results of which are shown in Table 2. Our indirect MMD losses improved the performance of repulsive MMD-GAN, and our KSD methods achieved the top performances. Thus, we chose MMD(δ) and KSD for ablation experiments.

4.2. Experimental Setup

Dataset. We compared the generative qualities of different losses based on CIFAR10 (50k training samples, 10 classes, 32^2 pixels) [17], CIFAR100 (50k training samples, 100 classes, 32^2 pixels) [17], and Tiny-ImageNet datasets (100k training samples, 200 classes, 64^2 pixels) [18]. Moreover, we trained DCGAN on the MNIST dataset (60k gray level samples, 10 classes, 28^2 pixels) [19] to demonstrate the distribution of discriminator outputs via tSNE [34] in Appendix and validate Section 3.4. We also trained the CelebA dataset (203k training samples, adjusted to 64^2 pixels) [25] and FFHQ dataset (7k images, 1024^2 pixels) [14] in Section 4.4.

Compared methods. We set three past methods as the baselines to compare with our approaches in Table 3. Firstly, the non-saturating loss function (Vanilla GAN) solved the saturation problem and showed its advantages compared with the original loss function [7, 6]. Next, we chose the well-known Wasserstein distance [1] with the hinge loss function [37] (Wasserstein-GAN) as another baseline method. The hinge loss has been validated in many past works and also achieved considerable success in BigGAN [2]. The third baseline method was the repulsive MMD loss [35] (Repulsive MMD-GAN). The repulsive MMD loss improved the performance of the original MMD-GAN [20]. We compared the generation qualities of four different settings of KSD-GAN, which were combinations of using Uniform or Normal distribution and using the hinged kernel for either the initial or all training steps.

Hyper-parameters. We used official PyTorch[31] implementation codes for training and evaluation. We only edited the last layer of the discriminator to multiple dimensions and used default settings as in BigGAN and StyleGAN2 (e.g. the learning rate, discriminator steps per generator step, and

Metrics	Generator loss	Discriminator loss
MMD	$\mathbb{E}[k(X, X')] + \mathbb{E}[k(Y, Y')]$ $-2 * \mathbb{E}[k(X, Y)]$ (Eq. 2)	$-\mathbb{E}[k(X, X')] - \mathbb{E}[k(Y, Y')]$ $+2 * \mathbb{E}[k(X, Y)]$ (negative Eq. 2)
MMD(Rep.)	As above	$\mathbb{E}[k(X, X')] - \mathbb{E}[k(Y, Y')]$ (Eq. 3)
MMD(δ)†	$\mathbb{E}[k(Y, Y')]$ $-2 * \mathbb{E}[k(0, Y)]$ (Eq. 8)	$\mathbb{E}[k(X, X')] - \mathbb{E}[k(Y, Y')]$ $-2 * \mathbb{E}[k(X, 0)] + 2 * \mathbb{E}[k(0, Y)]$ (Eq. 7)
MMD($\mathcal{N}$)†	$\mathbb{E}[k(Y, Y')]$ $-2 * \mathbb{E}[k(N, Y)]$	$\mathbb{E}[k(X, X')] - \mathbb{E}[k(Y, Y')]$ $-2 * \mathbb{E}[k(X, N)] + 2 * \mathbb{E}[k(N, Y)]$
KSD($\mathcal{U}$)†	$\mathbb{E}[u_{\mathcal{U}}(Y, Y')]$ (Eq. 9,$q = \mathcal{U}$)	$\mathbb{E}[u_{\mathcal{U}}(X, X')] - \mathbb{E}[u_{\mathcal{U}}(Y, Y')]$
KSD($\mathcal{N}$)†	$\mathbb{E}[u_{\mathcal{N}}(Y, Y')]$ (Eq. 9,$q = \mathcal{N}$)	$\mathbb{E}[u_{\mathcal{N}}(X, X')] - \mathbb{E}[u_{\mathcal{N}}(Y, Y')]$

Table 1: Equations in related works and our proposed loss metrics. '†': our proposals. 'Rep.': 'repulsive' in [35]. Here, we use 'X' and 'Y' to abbreviate '$\mathcal{D}(r), r \sim real$' and '$\mathcal{D}(g), g \sim fake$', respectively. '$N$': samples from the intermediate distributions $\mathcal{N}$. For KSD-based losses, we list cases of Uniform or Normal distribution as the intermediate distribution by substituting the q distribution in Eq. 9.

Metrics	Vanilla(JS)	Wasserstein	MMD(Rep.)	MMD(δ)	MMD($\mathcal{N}$)	KSD($\mathcal{U}$)	KSD($\mathcal{N}$)
FID↓	12.0797	15.2348	19.5360	12.6531	16.6766	**8.5005**	8.5883

Table 2: Results of preliminary StyleGAN2-based [15] unconditional generation experiments among different loss metrics. 'Rep.': 'Repulsive' and 'δ': the indirect MMD loss with Dirac-δ distribution. '$\mathcal{N}$' and '$\mathcal{U}$': intermediate distribution for the indirect adversarial losses.

betas in the Adam optimizer [16], etc.). Therefore our implementations were different from those original reports. For repulsive MMD-GAN experiments, we used the multi-scale RBF kernel with $\sigma \in \{1, \sqrt{2}, 2, 2\sqrt{2}, 4\}$, and our KSD used $\sigma = 1$.

Evaluation metrics. We used Inception score (IS, higher is better) [33], Fréchet Inception distance (FID, lower is better) [10], and Learned Perceptual Image Patch Similarity (LPIPS, higher is more diverse) [38] to validate the quantitative evaluations.

For each model, we sampled 50k samples randomly to calculate IS and FID. For LPIPS, we computed the score for every class in Table 4 to demonstrate the generation diversity.

4.3. Quantitative Analysis

Numerical Analysis As shown in Table 3, we compared the IS and FID for three real-world datasets. In all experiments, the training was performed with exactly the same settings except for the loss function. The results in Table 3 indicated the following: 1) First of all, our KSD-GAN has superior performance to the other loss functions. To be specific, KSD-GAN with a warm start and Uniform intermediate distribution shows its advantages in terms of FID; 2) Secondly, IS of KSD-GAN varies with the settings. CIFAR10 and CIFAR100 datasets contain images with relatively lower resolution, and Normal distribution performs higher than Uniform distribution in these cases; 3) Furthermore, the choice of the intermediate distribution affects the generation quality. Generally, Uniform distribution cases perform higher in terms of FID. We conjecture that Normal distribution cases tend to maintain a more concentrated feature space than the Uniform distribution, making it easier for the generator to cheat the discriminator while losing some details in the margin. On the other hand, Uniform distribution keeps the average prior information to acquire more details. Thus, in low-resolution learning cases, details have lower priority, and Normal distribution case is higher in terms of IS, while Uniform distribution is

Metrics	CIFAR10		CIFAR100		Tiny ImageNet	
	IS↑	FID↓	IS↑	FID↓	IS↑	FID↓
Vanilla (JS)	8.126±0.09	10.46	9.080±0.13	15.65	10.481±0.12	37.57
Wasserstein	7.554±0.09	10.99	9.272±0.15	12.07	12.666±0.16	23.83
MMD(Rep.)	7.396±0.06	10.17	6.513±0.11	30.43	6.080±0.07	74.13
MMD(δ)	9.279±0.08	7.70	9.577±0.16	8.83	10.943±0.12	28.76
KSD($\mathcal{N}$)-hinged	9.169±0.10	10.60	**10.312**±0.17	10.50	10.227±0.12	33.90
KSD($\mathcal{U}$)-hinged	9.166±0.08	7.15	9.831±0.18	7.63	**12.706**±0.18	21.48
KSD($\mathcal{N}$)-w.s.	**9.327**±0.10	12.52	9.736±0.13	11.87	10.715±0.16	28.43
KSD($\mathcal{U}$)-w.s.	9.128±0.09	**6.05**	9.781±0.06	**7.35**	12.205±0.22	**21.47**

Table 3: Results of BigGAN [2] conditional generation experiments. We compared four settings of our KSD-GAN with three past loss metrics. MMD(Rep.) stands for the repulsive MMD loss. In our methods, '-hinged' stands for using a hinged kernel as introduced in Appendix B. '-w.s.' refers to a warm start that applies a hinged kernel for first 10k iterations.

advantageous in higher resolution learning cases; 4) Besides, using a hinged kernel for the first 10k iterations yields FID results that are greater than IS. The IS results changed slightly, whereas the FID results improved slightly except for the Normal distribution case in the Tiny-ImageNet experiments. We speculate that removing the hinged kernel increases the volume of information for training, enabling the better discriminator, while also losing the protection of convergence, which leads to different performances; 5) In some cases, KSD-GAN with Normal distribution exhibited diverse performance, improving in one evaluation metric while becoming weaker in another. We show the generation diversity based on LPIPS to further analyze this phenomenon.

Diversity Analysis As shown in Table. 4, KSD-GAN experiments with Normal distribution generally showed a higher diversity than Uniform ones. We conjecture that Normal distribution cases tend to prioritize on generating high-quality images in several classes while neglecting details in other classes. Our MMD(δ) method shows a higher diversity than MMD(Rep.) in larger categories. We speculate that MMD(Rep.) may be trapped in trivial local minima and may sacrifice more details to minimize the losses before exploring a larger diversity, while ours can maintain more details in the optimization. Other statistics are available in Appendix.

Metrics (LPIPS↑)	CIFAR-10	CIFAR-100	Tiny-ImageNet
Vanilla(JS)	0.1923	0.1544	0.4872
Wasserstein	0.1815	0.2025	0.5092
MMD(Rep.)	0.1905	0.2061	0.4428
MMD(δ)	0.1781	**0.2273**	**0.6546**
KSD($\mathcal{U}$)	0.1723	0.1655	0.4841
KSD($\mathcal{N}$)	**0.1934**	0.1731	0.5519

Table 4: LPIPS mean values (higher is better) among different classes in three datasets.

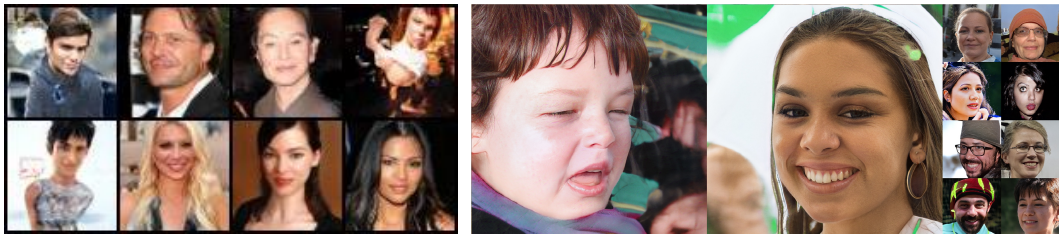
Metrics(FID↓)	CelebA	FFHQ
Vanilla(JS)	8.53	-
Wasserstein	7.13	7.41
MMD(Rep.)	12.78	-
MMD(δ)	3.94	5.33
KSD($\mathcal{U}$)	**3.63**	**4.82**

Table 5: FID results (lower is better) of unimodal dataset experiments among different loss metrics.

4.4. Qualitative Results on Human Face Generation

In our BigGAN experiments, the unimodal generation tasks were also trained on human face datasets

(a) KSD-GAN (CelebA) (b) KSD-GAN (FFHQ)



(c) Wasserstein (CelebA) (d) Wasserstein (FFHQ)

Figure 2: Human face image generation samples of our KSD-GAN with the Uniform intermediate distribution and the warm start settings compared with the Wasserstein baseline.

as shown in Fig. 2, and achieved FID of 3.63 for CelebA in Table 5. We also used different losses to continue training the pre-trained StyleGAN2 [13] and recorded their minimum FID score after they obtained a stable result. Our KSD losses achieved FID of 4.82 after continuedly learning on FFHQ. More samples are available in Appendix.

5. Limitations and Conclusion

We encountered several demerits, which were not perfectly solved in this work. Firstly, our methods follow other kernel-based methods, such as MMD-GAN, which have a comparatively more hyper-parameters to adjust. Our work mainly used similar hyper-parameters based on past literature while maintaining an additional tuning space. Secondly, our methods have higher computational complexity in the last layer of the discriminator than some other metrics. However, this would not be a problem at inferring stage. Finally, our methods may make it difficult to use existing pre-trained models. Traditional loss metrics have posteriorly defined outputs from the discriminator, while our proposals are towards one explicit intermediate distribution (as shown in Fig. 1).

In this work, our novel KSD loss function outperformed past loss functions in terms of IS and FID, and had comparable LPIPS scores for several real-world datasets. We may attempt to use an artificial intermediate distribution in the future for specific purposes. Owing to the simple prior hypothesis in Uniform and Normal distributions, moving the real or fake source close to the intermediate distribution is relatively straightforward. However, such existing distributions are unlikely to be the barycenter between real and fake distributions. Therefore, finding an analytical way to create the distribution as an intermediate distribution may improve the convergence speed or maintain more details during the training process. Moreover, using some pre-trained models as the intermediate distribution may achieve a particular domain transfer. Future works could explore the feasibility of these ideas.

Acknowledgement This work was supported by Institute of AI and Beyond of the University of Tokyo, JSPS/MEXT KAKENHI Grant Numbers 22K17947, JP19H04166 and JP22H05015.

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

[2] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2018.

[3] Imre Csiszár and Paul C Shields. Information theory and statistics: A tutorial. 2004.

[4] Wenbo Gong, Yingzhen Li, and José Miguel Hernández-Lobato. Sliced kernelized stein discrepancy. In *International Conference on Learning Representations*, 2020.

[5] Wenbo Gong, Kaibo Zhang, Yingzhen Li, and José Miguel Hernández-Lobato. Active slices for sliced stein discrepancy. In *International Conference on Machine Learning*, pages 3766–3776. PMLR, 2021.

[6] Ian Goodfellow. Nips 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*, 2016.

[7] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014.

[8] Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, and Richard Zemel. Learning the stein discrepancy for training and evaluating energy-based models without sampling. In *International Conference on Machine Learning*, pages 3732–3747. PMLR, 2020.

[9] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 5769–5779, 2017.

[10] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NIPS*, 2017.

[11] Tianyang Hu, Zixiang Chen, Hanxi Sun, Jincheng Bai, Mao Ye, and Guang Cheng. Stein neural sampler. *arXiv preprint arXiv:1810.03545*, 2018.

[12] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018.

[13] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *IEEE Conference on Neural Information Processing Systems;*, 2020.

[14] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

[15] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020.

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[17] Alex Krizhevsky et al. Learning multiple layers of features from tiny images. 2009.

[18] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7:7, 2015.

[19] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[20] Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabás Póczos. Mmd gan: Towards deeper understanding of moment matching network. In *NIPS*, 2017.

[21] Friedrich Liese and Igor Vajda. On divergences and informations in statistics and information theory. *IEEE Transactions on Information Theory*, 52(10):4394–4412, 2006.

[22] Qiang Liu, Jason Lee, and Michael Jordan. A kernelized stein discrepancy for goodness-of-fit tests. In *International conference on machine learning*, pages 276–284. PMLR, 2016.

[23] Qiang Liu and Dilin Wang. Stein variational gradient descent: a general purpose bayesian inference algorithm. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 2378–2386, 2016.

[24] Shuang Liu, Olivier Bousquet, and Kamalika Chaudhuri. Approximation and convergence properties of generative adversarial learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 5551–5559, 2017.

[25] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[26] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for gans. In *International Conference on Learning Representations*, 2018.

[27] Youssef Mroueh, Tom Sercu, and Vaibhava Goel. Mcgan: Mean and covariance feature matching gan. In *International conference on machine learning*, pages 2527–2535. PMLR, 2017.

[28] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyond. *arXiv preprint arXiv:1605.09522*, 2016.

[29] Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, pages 429–443, 1997.

[30] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. *Advances in neural information processing systems*, 29, 2016.

[31] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[32] Yunchen Pu, Zhe Gan, Ricardo Henao, Chunyuan Li, Shaobo Han, and Lawrence Carin. Vae learning via stein variational gradient descent. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 4239–4248, 2017.

[33] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *NIPS*, 2016.

[34] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[35] Wei Wang, Yuan Sun, and Saman Halgamuge. Improving mmd-gan training with repulsive loss function. In *International Conference on Learning Representations*, 2019.

[36] Jiasen Yang, Qiang Liu, Vinayak Rao, and Jennifer Neville. Goodness-of-fit testing for discrete distributions via stein discrepancy. In *ICML*, pages 5561–5570. PMLR, 2018.

[37] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International conference on machine learning*, pages 7354–7363. PMLR, 2019.

[38] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.