creative commons

C O M M O N S D E E D

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

Master's Thesis of Chaerin Kong

Data-efficient Generative Modeling

데이터 효율적 생성 모델

August 2023

Graduate School of Convergence Science and Technology
Seoul National University
Intelligence and Information Major

Chaerin Kong

Data Efficient Generative Modeling

Chaerin Kong

Submitting a master's thesis of
Intelligence and Information

August 2023

Graduate School of Convergence Science and Technology
Seoul National University
Intelligence and Information Major

Chaerin Kong

Confirming the master's thesis written by
Chaerin Kong
August 2023

Chair	박재흥	(Seal)
Vice Chair	곽노준	(Seal)
Examiner	전동석	(Seal)

Abstract

Generative models, such as GANs, typically need large-scale training with a substantial volume of images to generate a variety of realistic images. GANs trained on small dataset can easily memorize the training samples and display undesirable properties like stairlike latent space where interpolation in the latent space yields discontinuous transitions in the output image space. In this work, we consider a challenging task of pretraining-free few-shot image generation, and seek to train existing generative models while minimizing overfitting and mode collapse. We propose mixup-based distance regularization on the feature space of both the generator and the discriminator that encourages the two adversaries to reason not only about the scarce observed data points but the relative distances in the feature space they reside. Qualitative and quantitative evaluations on diverse datasets demonstrate that our method is generally applicable to existing models to improve both fidelity and diversity under few-shot setting.

Keyword : Generative Adversarial Networks, Few-shot Image Generation
Student Number : 2021-27642

Table of Contents

Chapter 1. Introduction

Figure 1: Cross-domain Correspondence adaptation of FFHQ source generator on various target domains (10-shot). We can see that finding a semantically similar source domain is crucial for CDC.

1.1. Study Background

Fascinating features of Generative Adversarial Networks (GANs) such as realistic sample quality and smooth semantic interpolation in the latent space have drawn huge attention from the research community, but what we have enjoyed with little gratitude claim their worth in the data-limited regime. As naive training of GANs with small datasets often results in suboptimal fidelity and diversity, many have proposed novel approaches specifically designed for few-shot image synthesis. Among the most successful are those adapting a generator pretrained on the source domain to the target domain [31, 34, 26] and those seeking generalization to unseen domains through feature fusion [16, 19]. Despite their impressive synthesis quality, these approaches are often critically constrained in practice as they all require semantic similarity between the large source domain dataset and the actual domain of interest [34], as illustrated in Fig. 1. For some domains like abstract art paintings, medical images and cartoon illustrations, it is very challenging to collect large scale dataset, while at the same time, finding an adequate source domain to transfer from is not straightforward either. To train GANs from scratch with limited data, several augmentation techniques [54, 22] and model architecture [27] have been proposed. Although these methods have delivered promising results on low-shot benchmarks consisting of hundreds to thousands of training images, they fall short for few-shot generation

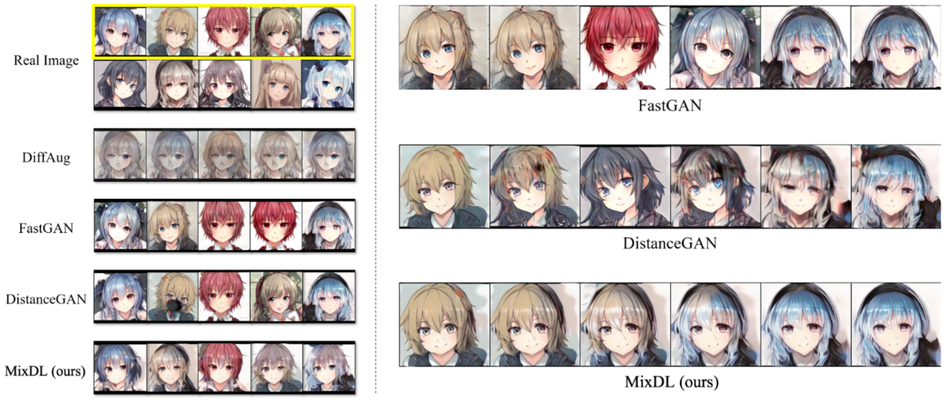

Figure 2: Training GANs with as little as 10 samples typically result in complete collapse or severe memorization (left). Strongly overfitted generators are only capable of producing a limited set of images, displaying stairlike latent interpolation (right).

where the dataset is even more constrained (e.g., n = 10).

GANs trained with small dataset typically display one of the two behaviors: severe quality degradation [54, 22] or complete loss of diversity [13], as visible from Fig. 2 (left). Therefore, producing a wide spectrum of good quality samples is the ultimate goal of few-shot generative models. We note that memorization differs from the classic mode collapse problem, as the former is not just lack of diversity, but the fundamental inability to generate unseen samples.

As directly mitigating memorization with as little as 10 training samples is extremely difficult if not impossible, we choose to tackle a surrogate problem instead. Our key observation is that strongly overfitted generators are only capable of producing a finite set of samples, resulting in discontinuous transitions in the image space upon latent interpolation. We name this stairlike latent space phenomenon, which has been pointed out by previous works [36, 8] as an indicator for memorization. Fig. 2 (right) demonstrates that previous methods designed for diversity preservation [4] or low-shot synthesis [27] all display such behavior under few-shot setting (n = 10). Therefore, instead of tackling the formidable task of suppressing memorization, we directly target stairlike latent space problem and propose effective distance regularizations to explicitly smooth the latent space of the generator (G) and the discriminator (D), which we empirically show is equivalent to fighting memorization in effect.

Our overall idea is to maximally exploit the scarce data points by continuously exploring their semantic mixups [51]. The discriminator overfitted to few real samples, however, displays overly confident and abrupt decision boundaries, leaving the generator with no choice but to faithfully reconstruct them in order to convince the opponent. This results in aforementioned stairlike latent space for both G and D, rendering smooth semantic mixups impossible. To overcome this challenge, we explore G's latent space with a randomly sampled interpolation coefficient c, enforcing relative semantic distances between samples to resemble the mixup ratio. By simultaneously imposing similar regularization on D's feature space, we prohibit the discriminator from embedding images to arbitrary locations for its convenience of memorizing, and guide its feature space to be aligned by semantic distances. Our objective is inspired by the formulation of [34] that aims to transfer diversity from the source domain to the target domain. We modify it for our single domain setting, where no source domain is available to import diversity from, and show that our method is capable of producing diverse novel samples with convincing quality even with as little as 10 training images. We further observe that models trained with our regularizations resist mode collapse surprisingly well even with no hand-designed augmentations. We believe that our distance regularizations encourage the model to preserve inherent diversity present in early stages throughout the course of training. Resistance to overfitting and mode collapse combined enables sample diversity under rigorous data constraint, which we demonstrate later with experimental results.

In sum, our contributions can be summarized as:

– We propose a two-sided distance regularization that encourages learning of smooth and mode-preserved latent space through controlled latent interpolation.

– We introduce a simple framework for few-shot image generation without a large source domain dataset that is compatible with existing architectures and augmentation techniques.

– We evaluate our approach on a wide range of datasets and demonstrate its effectiveness in generating diverse samples with convincing quality.

Chapter 2. Related Works

2.1. One-shot Image Generation

In order to create diverse outcomes from a single image, SinGAN [39] leverages the inherent ambiguity present in downsampled image. Based on SinGAN, ConSinGAN [18] proposes a technique to control the trade-off between fidelity and diversity. One-Shot GAN [41] uses a dual-branch discriminator where each head identifies context and layout, respectively. As one-shot image generation methods focus on exploiting a single image, they are not directly applicable to few-shot image generation tasks where the generator must learn the underlying distribution of a collection of images.

2.2. Low-shot Image Generation

Given a limited amount of training data, the discriminator in conventional GAN can easily overfit. To mitigate this problem, DiffAugment [54] imposes differentiable data augmentation to both real and fake samples while ADA [22] devises non-leaking adaptive discriminator augmentation. FastGAN [27] suggests a skip-layer excitation module and a self-supervised discriminator, which saves computational cost and stabilizes low-shot training. GenCo [11] shows impressive results on low-shot image generation task by using multiple discriminators to alleviate overfitting. Despite their promising performances on low-shot benchmarks, these methods often show significant instability under stricter data constraint, namely in few-shot setting.

2.3. Few-shot Generation with Auxiliary Dataset

Thus far, the few-shot image generation task ($n \approx 10$) mostly required pretraining on larger dataset with similar semantics [47, 46, 53, 37] mainly due to its inherent difficulty. A group of works [16, 19, 20, 3] learns transferable generation ability on seen categories and seek generalization into unseen categories through fusion-based methods. FreezeD [31] and EWC [26] further improves transfer learning framework for GANs. Meanwhile, CDC [34] computes the similarities between samples within each domain and encourages the corresponding similarity distributions to resemble each other. It aims to directly transfer the structural diversity of the source domain to the target, yielding impressive performance. In this paper, we modify the formulation of CDC and propose a novel few-shot generation framework that does not require any auxiliary data or separate pretraining step.

2.4. Generative Diversity

Mode collapse has been a long standing obstacle in GAN training. [2, 30] introduce divergence metrics that are effective at stabilizing GAN training while [12, 14] tackle this problem by training multiple networks. Another group of works [28, 29, 42, 49, 4] proposes regularization methods to preserve distances in the generated output space. Unlike these works, we consider the few-shot setting where the diversity is restricted mainly due to memorization, and introduce an interpolation-based distance regularization as an effective remedy.

2.5. Latent Mixup

Since [51], mixup methods have been actively explored to enforce smooth behaviors in between training samples [6, 43, 5]. In generative models, [36] emphasizes the importance of smooth latent transition as a counterevidence for memorization, but as state-of-the-art GAN models trained with sufficient data naturally possess such property [24, 8], it has been mainly studied with autoencoders. [7, 35] regularize autoencoders to learn smooth latent space while [48, 38] explore their potential as generative models through interpolation.

Chapter 3. Approach

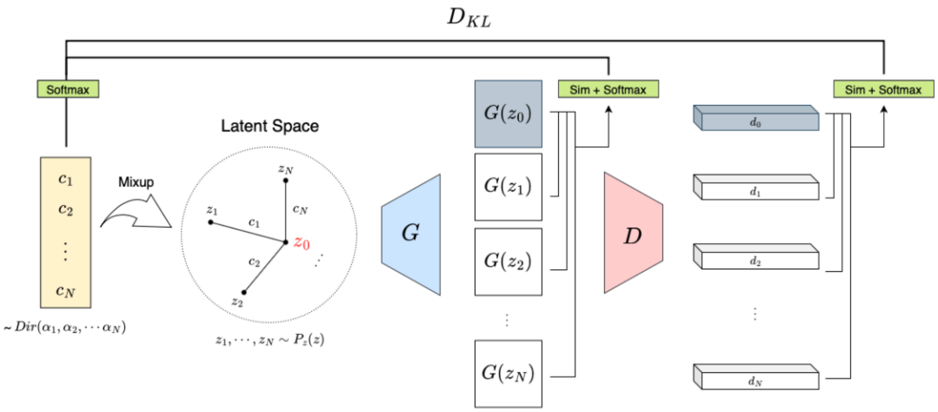

Figure 3: Overview of our Mixup-based Distance Learning (MixDL) framework.

We consider the situation where only few train examples (e.g., n = 10) are available with no semantically similar source domain. Hence, we would like to train a generative model from scratch, i.e., with no auxiliary dataset or separate pretraining step, using only a handful of images. Under such challenging constraints, overfitting greatly restricts a model's ability to learn data distribution and produce diverse samples. We identify its byproduct stairlike latent space as the core obstacle, as it not only indicates memorizing but also prohibits hallucination through semantic mixup. We observe that both the generator and the discriminator suffer from the problem with insufficient data, evidenced by discontinuous latent interpolation and overly confident decision boundary, respectively.

To this end, we propose mixup-based distance learning (MixDL) framework that guides the two players to form soft latent space and leverage it to generate diverse samples. We further discover that our proposed regularizers effectively combat mode collapse, a problem particularly more devastating with a small dataset, by preserving diversity present in early training stages. As our formulation is inspired by [34], we first introduce their approach in Sec. 3.1, and formally state our methods in Sec. 3.2 and Sec. 3.3. Our final learning framework and the corresponding details can be found in Sec. 3.4.

3.1. Cross-domain Correspondence

In CDC [34], the authors propose to transfer the relationship learned in a source domain to a target domain. They define a probability distribution from pairwise similarities of generated samples in both domains and bind the latter to the former. Formally, they define distributions as

$$p^l = softmax(\{sim(G_s^l(z_0), G_s^l(z_i))\}_{i=1}^N)$$

$$q^l = softmax(\{sim(G_{s\to t}^l(z_0), G_{s\to t}^l(z_i))\}_{i=1}^N)$$

where G_l is the generator activation at the l-th layer and $\{z_i\}_0^N$ are latent vectors. Note that G_s and $G_{s\to t}$ correspond to source and target domain generator, respectively, and p_l , q_l are N-way discrete probability distributions consisting of N pairwise similarities. Then, along with adversarial objective L_{adv}, they impose a KL-divergence-based regularization of the following form:

$$L_{dist} = E_{z\sim p_z(z)}[D_{KL}(q^l||p^l)]$$

The benefits of this auxiliary objective are twofold: it prevents distance collapse in the target domain and transfers diversity from the source to target via one-to-one correspondence. However, as visible from Fig. 1, the synthesis quality is greatly affected by the semantic distance between source and target. Hence, we propose MixDL, which modifies CDC for pretraining-free few-shot image synthesis and provides consistent performance gains across different benchmarks.

3.2. Generator Latent Mixup

In [34], the anchor point z0 could be chosen arbitrarily from the prior distribution $p_z(z)$ since they were transferring the rich structural diversity of the source domain to the target latent space. As this is no longer applicable in our setting, we propose to resort to diverse combinations of given samples. Hence, preserving the modes and learning interpolable latent space are our two main desiderata. To this end, we define our anchor point using Dirichlet distribution as follows:

$$z_0 = \Sigma_{i=1}^N c_i z_i \, , c \sim Dir(\alpha_1, \dots, \alpha_N)$$

where $c = [c_1, c_2, \dots, c_N]^T$. Using the above equation, the latent space can be navigated in a quantitatively controlled manner. Defining probability distribution of pairwise similarities as in [34], we bind it to the interpolation coefficients c instead. The proposed distance loss is defined as follows:

$$L_{dist}^G = E_{z\sim p_z(z), c\sim Dir(\alpha)}[D_{KL}(q^l||p)]$$

$$q^l = softmax(\{sim(G^l(z_0), G^l(z_i))\}_{i=1}^N)$$

$$p = softmax(\{c_i\}_{i=1}^N)$$

where Dir(α) denotes the Dirichlet distribution with parameters α = $(\alpha_1, \ldots, \alpha_N)$. This efficiently accomplishes our desiderata. Intuitively, unlike naive generators that gradually converge to few modes, our regularization forces the generated samples to differ from each other by a controlled amount, making mode collapse very difficult. At the same time, we constantly explore our latent space with continuous coefficient vector c, explicitly enforcing smooth latent interpolation.
An anchor point similar to [34] can be obtained with one-hot coefficients c.

3.3. Discriminator Feature Space Alignment

While the generator distance regularization can alleviate mode collapse and stairlike latent space problem surprisingly well, the root cause of constrained diversity still remains unresolved, i.e., discriminator overfitting. As long as the discriminator delivers overconfident gradient signals to the generator based on few examples it observes, generator outputs will be strongly pulled towards the small set of observed data. To encourage the discriminator to provide smooth signals to the generator based on reasoning about continuous semantic distances rather than simply memorizing the data points, we impose similar regularization on its feature space. Formally, we define our discriminator D(x) = $(d^2 \circ d^1)$(x) where d^2(x) refers to the final FC layer that outputs {real, fake}. When a set of generated samples $\{G(z_i)\}_{i=1}^N$ and the interpolated sample $G(z_0)$ is provided to D, we construct an N-way distribution similar to the above equations as

$$r = softmax(\{sim(proj(d_0^1), proj(d_i^1))\}_{i=1}^N)$$

where proj refers to a linear projection layer widely used in self-supervised learning literature [9, 10, 15] and $d_j^1 = d^1(G(z_j))$. Without the linear projector, we found the constraint too rigid that it harms overall output quality. We define our distance regularization for the discriminator as

$$L_{dist}^d = E_{z \sim p_z(z), c \sim Dir(\alpha)}[D_{KL}(r || p)]$$

This regularization penalizes the discriminator for storing memorized real samples in arbitrary locations in the feature space and encourages the space to be aligned with relative semantic distances. Thus it makes memorization harder while guiding the discriminator to provide smoother and more semantically meaningful signals to the generator.

3.4. Final Objectives

Fig. 3 shows an overall concept of our method. Our final objective takes the form:

$$L^G = L^G_{adv} + \lambda_G L^G_{dist}$$

$$L^D = L^D_{adv} + \lambda_D L^D_{dist}$$

where we generally set $\lambda_G = 1000$ and $\lambda_D = 1$.

As our method is largely independent of model architectures, we apply our method to two existing models, StyleGAN24[24] and FastGAN[27]. We keep their objective functions as they are and simply add our regularization terms. For StyleGAN2, we interpolate in W rather than Z, which has been shown to have better properties such as disentanglement [44, 55, 1]. Mixup coefficients c is sampled from a Dirichlet distribution of parameters all equal to one. Patch-level discrimination [21, 34] is applied for mixup images to encourage our generator to be creative while exploring the latent space.

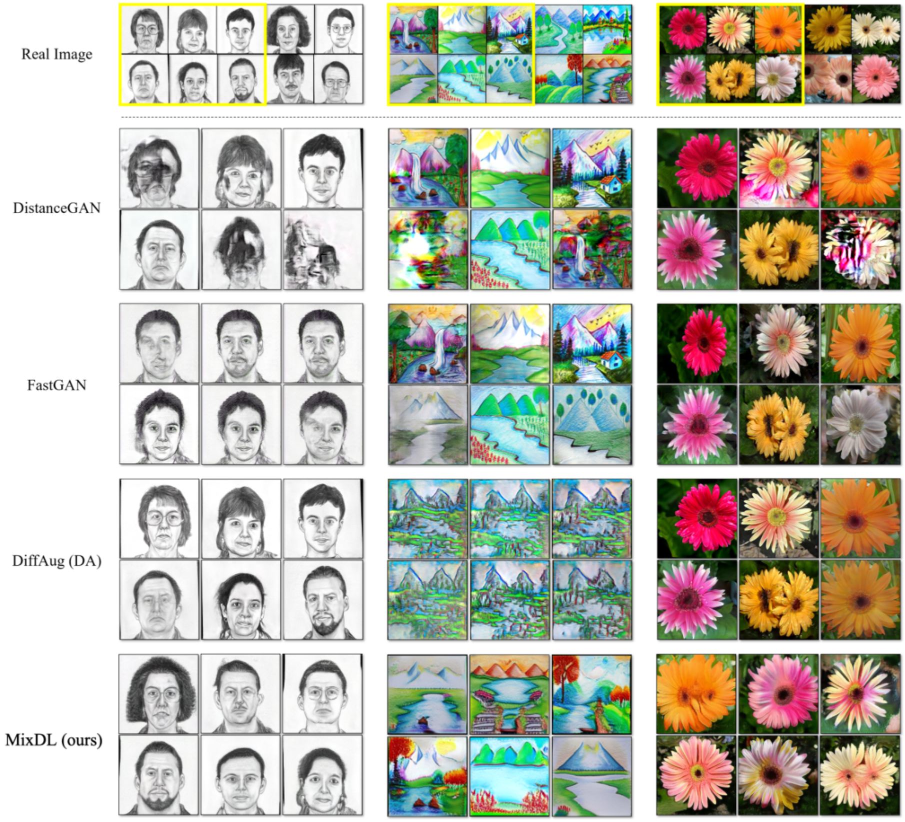

Figure 4: 10-shot image generation results. While baseline methods either collapse or simply replicate the training samples (yellow box), our method actively encourages the generator to explore semantic mixups of given samples, which enables synthesis of various unseen samples.

We mainly apply our method to the state-of-the-art unconditional GAN model, StyleGAN2 [24]. Data augmentation techniques introduced by [54] and [22] show promising performance on low-shot image generation task, so we evaluate them along with ours and refer to them as DA and ADA respectively. We additionally apply our method to FastGAN [27], which is a light-weight GAN architecture that allows faster convergence with limited data. Although methods designed for alleviating mode collapse [4, 28, 29] are not directly targeted for data-limited setting, we further adopt these as baselines considering the similarity in objective formulation. We implement them on StyleGAN2 for better synthesis quality and fair comparison. Transfer based methods such as EWC [26] and CDC [34] fundamentally differ from ours as they require a large scale pre training and thus are not directly comparable. However, we include CDC [34] since our method adjusts

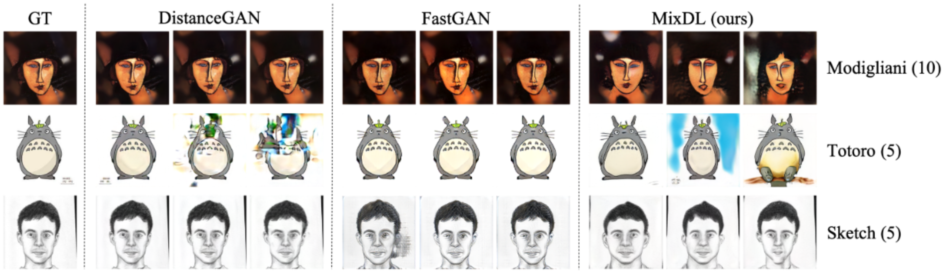

Figure 5: Uncurated collection of samples sharing the same training image as nearest neighbor.

it for a more general single domain setting.

For quantitative evaluation, we use Animal-Face Dog [40], Oxford-flowers [33], FFHQ-babies [23], face sketches [45], Obama and Grumpy Cat [54], anime face [27] and Pokemon (pokemon.com, [27]). Aforementioned datasets contain 100 to 8189 samples, so we simulate few-shot setting by randomly sampling 10 images, if not stated otherwise. For qualitative evaluation, we further experiment on paintings of Amedeo Modigliani [50], landscape drawings [34] and web-crawled images of Totoro. All the images are 256×256. Additional synthesis results and information about datasets can be found in the supplementary.

We measure Fr´echet Inception Distance (FID) [17], sFID [32] and precision/recall [55] for datasets containing a sufficient number ($\geq$ 100) of samples along with pairwise Learned Perceptual Image Patch Similarity (LPIPS) [52]. For simulated few-shot tasks, the FID and sFID are computed against the full dataset as in [26, 34]. We further use LPIPS as a distance metric for demonstrating interpolation smoothness and mode preservation.

4.1. Qualitative Results

Fig. 4 shows generated samples from 10-shot training. We observe that baseline methods either collapse to few modes or severely overfit to the training data, resulting in inability to generate novel samples. Ours is the only method that produces a variety of convincing samples that are not present in the training set. Our method combines visual attributes such as hairstyle, beard and glasses in a natural way, producing distinctive samples under harsh data constraint.

The difference is more distinguished when we take a closer look. In Fig. 5 we display uncurated sets of generated images along with their nearest neighbor real images. Samples from DistanceGAN [4] and FastGAN [27] are either defective or largely identical to the

Table 1: Quantitative results on 10-shot image generation benchmark.

Method	Anime Face			Animal-Face Dog			Oxford Flowers			Face Sketches			Pokemon		
	FID ↓	sFID ↓	LPIPS ↑	FID ↓	sFID ↓	LPIPS ↑	FID ↓	sFID ↓	LPIPS ↑	FID ↓	sFID ↓	LPIPS ↑	FID ↓	sFID ↓	LPIPS ↑
FastGAN [27]	123.7	127.9	0.341	103.0	117.4	0.633	182.7	111.2	0.667	76.3	81.8	0.148	123.5	105.7	0.578
StyleGAN2 [23]	166.0	111.4	0.363	177.5	127.7	0.569	177.3	143.0	0.537	94.2	84.4	0.435	257.6	136.5	0.439
StyleGAN2 + DA [54]	162.0	96.8	0.204	136.1	123.5	0.559	187.0	154.4	0.687	43.1	59.9	0.438	280.1	148.9	0.179
StyleGAN2 + ADA [22]	130.2	108.0	0.288	236.5	126.2	0.636	167.8	83.5	0.719	62.8	67.3	0.399	214.3	95.5	0.496
FastGAN + Ours	107.6	98.5	0.478	99.8	111.7	0.625	180.5	75.5	0.657	45.0	58.0	0.416	144.0	118.3	0.584
StyleGAN2 + Ours	73.1	**92.8**	0.548	96.0	99.9	0.682	136.6	67.6	0.734	39.4	**43.3**	0.479	117.0	**57.7**	0.539
StyleGAN2 + DA + Ours	**70.2**	94.1	0.551	96.4	107.6	0.682	129.9	66.9	0.705	**35.6**	50.1	0.471	**114.3**	79.0	**0.607**
StyleGAN2 + ADA + Ours	75.0	96.5	**0.571**	**94.1**	**96.6**	**0.684**	**127.7**	**52.5**	**0.763**	39.2	45.7	**0.482**	155.5	65.7	0.544
StyleGAN2 + CDC† [34]	93.4	107.4	0.469	206.7	110.1	0.545	107.5	99.9	0.518	45.7	46.1	0.428	126.6	79.1	0.342

Table 2: Quantitative comparison with diversity preservation methods on 10-shot tasks.

Method	Anime Face			Animal-Face Dog			FFHQ-babies		
	FID ↓	sFID ↓	LPIPS ↑	FID ↓	sFID ↓	LPIPS ↑	FID ↓	sFID ↓	LPIPS ↑
N-Div [28]	175.4	176.4	0.425	150.4	153.6	0.632	177.1	177.1	0.510
MSGAN [29]	138.6	100.5	0.536	165.7	123.0	0.630	165.4	120.1	0.569
DistanceGAN [4]	84.1	93.0	0.543	102.6	114.2	0.678	105.7	102.9	0.640
MixDL (ours)	**73.1**	**92.8**	**0.548**	**96.0**	**99.9**	**0.682**	**83.4**	**73.9**	**0.643**

corresponding GT, but our method generates unique samples with recognizable visual features. We believe this is because our distance regularization enforces outputs from different latent vectors to differ from each other, proportionally to the relative distances in the latent space.

4.2. Quantitative Evaluation

Tab. 1 shows FID, sFID and LPIPS scores for several low-shot generation methods [54, 22, 27] on 10-shot image generation task. We can see that our method consistently outperforms the baselines, often with significant margins. Moreover, our regularizations can be applied concurrently to data augmentations to obtain further performance gains. Note that while StyleGAN2 armed with advanced data augmentations fails to converge from time to time, our method guarantees stable convergence to a better optimum across all datasets. Surprisingly, ours outperforms CDC [34] on all metrics even when the two domains are closely related, e.g. anime-face and face sketches. For dissimilar domains like pokemon, CDC tends to sacrifice diversity (i.e., LPIPS) for better fidelity, which nevertheless falls short overall. We present training snapshots in the supplementary.

Additional quantitative comparison with diversity preserving methods is displayed in Tab. 2. Although these methods have some similarities with ours, especially MixDL-G, we can observe steady improvements with MixDL. As the baselines are simply designed to

Table 3: FID-comparison.

Dataset	Obama	Cat	Flowers	Obama	Cat
Shot	100	100	100	10	10
LPIPS	0.615	0.613	0.795	0.598	0.598
StyleGAN2	63.1	43.3	192.2	174.7	76.4
+ DA	46.9	27.1	91.6	66.8	45.6
+ Ours	58.4	26.6	82.0	62.7	41.1
+ DA + Ours	**45.4**	**26.5**	**64.0**	**57.9**	**39.3**

Table 4: Precision and recall.

Method	Obama		Cat	
	Prec.	Rec.	Prec.	Rec.
StyleGAN2	0.47	0.07	0.15	0.12
+MixDL	**0.52**	**0.32**	**0.86**	**0.50**
FastGAN	0.90	0.36	0.90	0.43
+MixDL	**0.91**	**0.47**	**0.91**	**0.50**

Table 5: Ablation on loss components.

MixDL		Dog (10-shot)		Babies (100-shot)		Flowers (100-shot)	
G	*D*	FID ↓	LPIPS ↑	FID ↓	LPIPS ↑	FID ↓	LPIPS ↑
		177.5	0.569	131.0	0.574	192.2	0.747
	✓	118.4	0.649	83.4	0.638	94.1	0.775
✓		**95.4**	0.673	71.7	0.638	84.0	0.780
✓	✓	96.0	**0.682**	**63.4**	**0.647**	**82.0**	**0.782**

Table 6: Coefficient sampling.

Distribution	Dirichlet			Gaussian	Uniform
Parameter	$\alpha = 0.1$	$\alpha = 1$	$\alpha = 10$	standard	-
FID (↓)	76.4	**73.1**	80.8	76.0	74.8
LPIPS (↑)	0.536	**0.548**	0.532	**0.548**	0.546

minimize mode collapse, we believe they are relatively prone to memorization, which is a far devastating issue in few-shot setting.

While pretraining-free 10-shot image synthesis task has not been studied much, several works [27, 54] have previously explored generative modeling with as little as 100 samples. We present quantitative evaluations on popular low-shot benchmarks in Table 3. We observe that our method consistently improves the baseline, and the margin is larger for more challenging tasks, i.e., dataset with greater diversity or fewer training samples. We discuss experiments on these benchmarks in depth in Sec. 5. Tab. 4 shows precision and recall [25] for these benchmarks, where MixDL boosts scores especially in terms of diversity.

4.3. Ablation Study

We further evaluate the effects of the proposed regularizations, MixDL-G (generator) and MixDL-D (discriminator), through ablation under different settings. In Tab. 5, we observe that in general, our regularizations both contribute to better quality and diversity, while in some special cases, only adding MixDL-G leads to better FID score. We conjecture that aligning discriminator's feature vectors with the interpolation coefficients can impose overly strict constraint for some datasets. We nonetheless observe consistent improvements on diversity.

Fig. 6 shows the evaluation across different subset sizes. Since FFHQ-babies and Oxford-flowers contain more than 2,000 and 8,000 images respectively, we randomly sample subsets of size 10, 100 and 1,000. We can see that the performance of StyleGAN2 steadily improves with more training samples, but it consistently benefits from

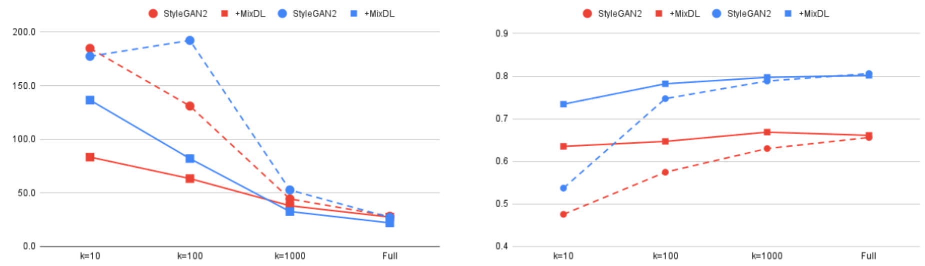

Figure 6: Shot ablation results. Red indicates FFHQ-babies and blue represents flowers. (left) FID scores (right) LPIPS score.

MixDL. Hence, we believe that with limited data in general, our method can be broadly used to improve model performance. Lastly in Tab. 6, the effect of using different Dirichlet concentration parameters and sampling distribution for mixup is illustrated. We find that setting α = 1 yields the best performance, so we uniformly use this throughout the experiments.

4.4. Latent Space Smoothness

Smooth latent space interpolation is an important property of generative models that disproves overfitting and allows synthesis of novel data samples. As our proposed method focuses on diversity through latent smoothing, we quantitatively evaluate this using a variant of Perceptual Path Length (PPL) proposed by [23].

PPL was originally introduced as a measure of latent space disentanglement under the assumption that a more disentangled latent space would show smoother interpolation behavior [23]. As we wish to directly quantify latent space smoothness, we slightly modify the metric by taking 10 subintervals between any two latent vectors and measure their perceptual distances. Tab. 7 reports the subinterval mean, standard deviation, and the mean for the full path (End). Note that as PPL is a quadratic measure, the sum of subinterval means can be smaller than the endpoint mean. All four models show similar endpoint mean, suggesting that the overall total perceptual distance is consistent, while ours displays the lowest PPL standard deviation. As low PPL variance across subintervals is a direct sign of perceptually uniform latent transitions, we can verify the effectiveness of our method in smoothing the latent space. Similar insight can be found from Fig. 7 where the baselines display stairlike latent transition while ours shows smooth semantic interpolation. More details on PPL computation can be found in the supplementary materials.

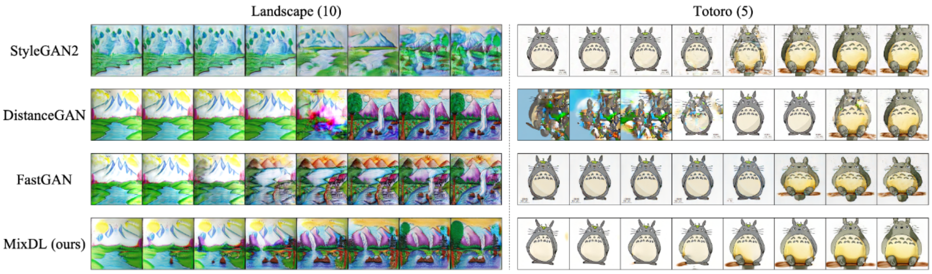

Figure 7: Latent space interpolation results.

Table 7: Perceptual Path Length Uniformity.

Dataset	Landscape			Totoro		
Metric	Mean	Std.	End	Mean	Std.	End
StyleGAN2	21.91	12.66	60.90	16.43	15.39	56.53
DistanceGAN	23.07	21.53	70.71	16.76	14.82	61.50
FastGAN	15.49	15.00	67.75	10.03	12.14	54.16
MixDL	12.82	**4.19**	64.28	11.75	**6.44**	56.83

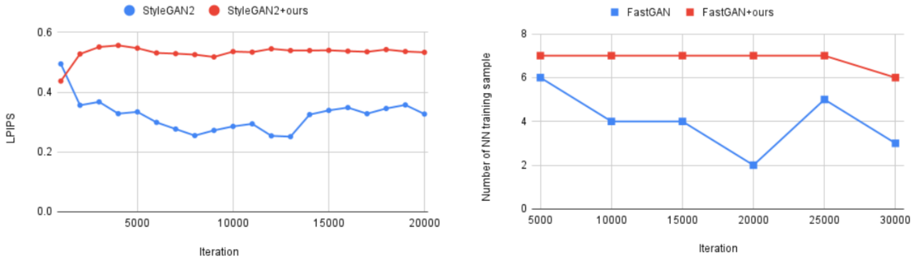

Figure 8: (left) LPIPS in early iterations (right) number of unique NN training samples.

4.5. Preserving Diversity

As opposed to [34] that preserves diversity in the source domain, our method can be interpreted as preserving the diversity inherently present in the early stages throughout the course of training, by constantly exploring the latent space and enforcing relative similarity between samples. To validate our hypothesis, we keep track of pairwise LPIPS of generated samples and the number of modes in the early iterations. Fig. 8 shows the result, where the number of modes is represented by the number of unique training samples (real images) that are the nearest neighbor to any of the generated images. In Fig. 8a, we can see that vanilla StyleGAN2 and our method show similar LPIPS in the beginning, but the baseline quickly loses diversity while ours maintains relatively high level of diversity throughout the training.

Fig. 8b delivers similar implication that FastGAN trained with MixDL better preserves modes compared to the baseline.

Combined with latent space smoothness explained in Sec. 4.4, generators equipped with MixDL learn rich mode-preserving latent space with smooth interpolable landscape. This naturally allows generative diversity particularly appreciated under the constraint of extremely limited data.

Chapter 5. Discussion

The trade-off between fidelity and diversity in GANs has been noted by many [8, 23]. Truncation trick, a technique widely used in generative models, essentially denotes that diversity can be traded for fidelity. In few-shot generation task, it is very straightforward to obtain near-perfect fidelity at the expense of diversity as one can simply overfit the model, while generating diverse unseen data points is very challenging. This implies that with only a handful of data, the diversity should be credited no less than the fidelity.

However, we believe that the widely used low-shot benchmarks, e.g., 100-shot Obama and Grumpy Cat, inherently favor faithful reconstruction over audacious exploration. The main limitations we find in these datasets are twofold: (i) the intra-diversity is too limited as they contain photos of a single person or object, evidenced by low LPIPS in Tab. 3 and (ii) FID is computed based on the 100 samples that were used for training. We acknowledge that (ii) is a common practice in generative models, but the problem with these benchmarks is that the number of samples is too limited, making it possible for some models to simply memorize a large portion of them. These two combined results in benchmarks that allow relatively easy replication and reward it generously at the same time. In other words, we believe that a model's capacity to explore continuous image manifold and be creative can potentially backfire in these benchmarks.

To address these limitations, in Tab. 3 we extend the benchmark with three additional datasets: 100-shot flowers, 10-shot Obama and Grumpy Cat. The first one challenges the model with greater diversity while the last two evaluate its capacity to learn distribution in a generalizable manner, as the FID is still computed against the full 100 images. As our method mainly aims for modeling diversity, we observe marginal performance gains in the traditional benchmarks. However on the extended benchmarks, it shows significant contributions, confirming its excellence at learning diversity even under challenging situations.

Chapter 6. Conclusion

We propose MixDL, a set of distance regularizations that can be directly added to existing models for few-shot image generation. Unlike previous works, MixDL enables high-quality synthesis of novel images with as few as 5 to 10 training samples, even without any source domain pretraining. Thorough evaluations on diverse benchmarks demonstrate the effectiveness of our framework. We hope our work facilitates future research on data efficient generative modeling.

Bibliography

1. Alaluf, Y., Patashnik, O., Cohen-Or, D.: Restyle: A residual-based stylegan encoder via iterative refinement. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6711-6720 (2021)
2. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: International conference on machine learning. pp. 214–223. PMLR (2017)
3. Bartunov, S., Vetrov, D.: Few-shot generative modelling with generative matching networks. In: International Conference on Artificial Intelligence and Statistics. pp. 670–678. PMLR (2018)
4. Benaim, S., Wolf, L.: One-sided unsupervised domain mapping. Advances in neural information processing systems 30 (2017)
5. Berthelot, D., Carlini, N., Cubuk, E.D., Kurakin, A., Sohn, K., Zhang, H., Raffel, C.: Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. arXiv preprint arXiv:1911.09785 (2019)
6. Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., Raffel, C.: Mixmatch: A holistic approach to semi-supervised learning. arXiv preprint arXiv:1905.02249 (2019)
7. Berthelot, D., Raffel, C., Roy, A., Goodfellow, I.: Understanding and improving interpolation in autoencoders via an adversarial regularizer. arXiv preprint arXiv:1807.07543 (2018)
8. Brock, A., Donahue, J., Simonyan, K.: Large scale gan training for high fidelity natural image synthesis. arXiv preprint arXiv:1809.11096 (2018)
9. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
10. Chen, X., He, K.: Exploring simple siamese representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15750–15758 (2021)

11. Cui, K., Huang, J., Luo, Z., Zhang, G., Zhan, F., Lu, S.: Genco: Generative co-training on data-limited image generation. arXiv preprint arXiv:2110.01254 (2021)
12. Durugkar, I., Gemp, I., Mahadevan, S.: Generative multi-adversarial networks. arXiv preprint arXiv:1611.01673 (2016)
13. Feng, Q., Guo, C., Benitez-Quiroz, F., Martinez, A.M.: When do gans replicate? on the choice of dataset size. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6701–6710 (2021)
14. Ghosh, A., Kulharia, V., Namboodiri, V.P., Torr, P.H., Dokania, P.K.: Multi-agent diverse generative adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8513–8521 (2018)
15. Grill, J.B., Strub, F., Altch´e, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Pires, B.A., Guo, Z.D., Azar, M.G., et al.: Bootstrap your own latent: A new approach to self-supervised learning. arXiv preprint arXiv:2006.07733 (2020)
16. Gu, Z., Li, W., Huo, J., Wang, L., Gao, Y.: Lofgan: Fusing local representations for few-shot image generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8463–8471 (2021)
17. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems 30 (2017)
18. Hinz, T., Fisher, M., Wang, O., Wermter, S.: Improved techniques for training single-image gans. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1300–1309 (2021)
19. Hong, Y., Niu, L., Zhang, J., Zhang, L.: Matchinggan: Matching-based few-shot image generation. In: 2020 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (2020)
20. Hong, Y., Niu, L., Zhang, J., Zhao, W., Fu, C., Zhang, L.: F2gan: Fusing-and-filling gan for few-shot image generation. In: Proceedings of the 28th ACM International Conference on

Multimedia. pp. 2535–2543 (2020)
21. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017)
22. Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., Aila, T.: Training generative adversarial networks with limited data. arXiv preprint arXiv:2006.06676 (2020)
23. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4401–4410 (2019)
24. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8110–8119 (2020)
25. Kynk ̈a ̈anniemi, T., Karras, T., Laine, S., Lehtinen, J., Aila, T.: Improved precision and recall metric for assessing generative models. Advances in Neural Information Processing Systems 32 (2019)
26. Li, Y., Zhang, R., Lu, J., Shechtman, E.: Few-shot image generation with elastic weight consolidation. arXiv preprint arXiv:2012.02780 (2020)
27. Liu, B., Zhu, Y., Song, K., Elgammal, A.: Towards faster and stabilized gan training for high-fidelity few-shot image synthesis. In: International Conference on Learning Representations (2020)
28. Liu, S., Zhang, X., Wangni, J., Shi, J.: Normalized diversification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10306–10315 (2019)
29. Mao, Q., Lee, H.Y., Tseng, H.Y., Ma, S., Yang, M.H.: Mode seeking generative adversarial networks for diverse image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1429–1437 (2019)
30. Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Paul Smolley, S.: Least squares generative adversarial networks. In: Proceedings of

the IEEE international conference on computer vision. pp. 2794–2802 (2017)
31. Mo, S., Cho, M., Shin, J.: Freeze the discriminator: a simple baseline for fine-tuning gans. arXiv preprint arXiv:2002.10964 (2020)
32. Nash, C., Menick, J., Dieleman, S., Battaglia, P.W.: Generating images with sparse representations. arXiv preprint arXiv:2103.03841 (2021)
33. Nilsback, M.E., Zisserman, A.: A visual vocabulary for flower classification. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). vol. 2, pp. 1447–1454. IEEE (2006)
34. Ojha, U., Li, Y., Lu, J., Efros, A.A., Lee, Y.J., Shechtman, E., Zhang, R.: Few-shot image generation via cross-domain correspondence. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10743–10752 (2021)
35. Oring, A., Yakhini, Z., Hel-Or, Y.: Autoencoder image interpolation by shaping the latent space. arXiv preprint arXiv:2008.01487 (2020)
36. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434 (2015)
37. Robb, E., Chu, W.S., Kumar, A., Huang, J.B.: Few-shot adaptation of generative adversarial networks. arXiv preprint arXiv:2010.11943 (2020)
38. Sainburg, T., Thielk, M., Theilman, B., Migliori, B., Gentner, T.: Generative adversarial interpolative autoencoding: adversarial training on latent space interpolations encourage convex latent distributions. arXiv preprint arXiv:1807.06650 (2018)
39. Shaham, T.R., Dekel, T., Michaeli, T.: Singan: Learning a generative model from a single natural image. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4570–4580 (2019)
40. Si, Z., Zhu, S.C.: Learning hybrid image templates (hit) by information projection. IEEE Transactions on pattern analysis and

machine intelligence 34(7), 1354–1367 (2011)
41. Sushko, V., Gall, J., Khoreva, A.: One-shot gan: Learning to generate samples from single images and videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2596–2600 (2021)
42. Tran, N.T., Bui, T.A., Cheung, N.M.: Dist-gan: An improved gan using distance constraints. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 370–385 (2018)
43. Verma, V., Kawaguchi, K., Lamb, A., Kannala, J., Solin, A., Bengio, Y., Lopez-Paz, D.: Interpolation consistency training for semi-supervised learning. Neural Networks (2021)
44. Wang, T., Zhang, Y., Fan, Y., Wang, J., Chen, Q.: High-fidelity gan inversion for image attribute editing. arXiv preprint arXiv:2109.06590 (2021)
45. Wang, X., Tang, X.: Face photo-sketch synthesis and recognition. IEEE transactions on pattern analysis and machine intelligence 31(11), 1955–1967 (2008)
46. Wang, Y., Gonzalez-Garcia, A., Berga, D., Herranz, L., Khan, F.S., Weijer, J.v.d.: Minegan: effective knowledge transfer from gans to target domains with few images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9332–9341 (2020)
47. Wang, Y., Wu, C., Herranz, L., van de Weijer, J., Gonzalez-Garcia, A., Raducanu, B.: Transferring gans: generating images from limited data. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 218–234 (2018)
48. Wertheimer, D., Poursaeed, O., Hariharan, B.: Augmentation-interpolative autoencoders for unsupervised few-shot image generation. arXiv preprint arXiv:2011.13026 (2020)
49. Yang, D., Hong, S., Jang, Y., Zhao, T., Lee, H.: Diversity-sensitive conditional generative adversarial networks. arXiv preprint arXiv:1901.09024 (2019)
50. Yaniv, J., Newman, Y., Shamir, A.: The face of art: landmark detection and geometric style in portraits. ACM Transactions on graphics (TOG) 38(4), 1–15 (2019)

51. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017)
52. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
53. Zhao, M., Cong, Y., Carin, L.: On leveraging pretrained gans for generation with limited data. In: International Conference on Machine Learning. pp. 11340–11351. PMLR (2020)
54. Zhao, S., Liu, Z., Lin, J., Zhu, J.Y., Han, S.: Differentiable augmentation for data-efficient gan training. arXiv preprint arXiv:2006.10738 (2020)
55. Zhu, P., Abdal, R., Qin, Y., Femiani, J., Wonka, P.: Improved stylegan embedding: Where are the good latents? arXiv preprint arXiv:2012.09036 (2020)

Abstract

생성 모델인 GAN과 같은 도구를 사용하여 다양하고 현실적인 이미지를 생성하는 것은 대량의 이미지로 대규모 훈련을 요구하는 경우가 일반적이다. 제한된 데이터로 훈련된 GAN들은 손쉽게 적은 수의 훈련 샘플들을 기억하게 되고, 출력 공간에서 불연속적인 전환을 초래하는 계단형 잠재 공간과 같은 불필요한 특성들을 나타내게 된다. 본 연구에서는 사전훈련이 없는 소량샘플 이미지 합성이라는 어려운 과제를 고려하고, 최소한의 과적합과 모드 붕괴로 기존 생성 모델을 훈련하는 방법을 탐구한다. 우리는 생성자와 상대판별자의 특성 공간에 대한 믹스업 기반 거리 정규화를 제안하며, 이것은 두 플레이어가 희박한 관측 데이터 포인트뿐만 아니라 그들이 존재하는 특성 공간에서의 상대적 거리에 대해 이해하도록 장려한다. 다양한 데이터셋에서의 질적 및 양적 평가 결과가 우리의 방법이 소량샘플 설정 하에서 생성 이미지의 품질과 다양성을 모두 향상시킬 수 있음을 보여준다.