Smoothing the Generative Latent Space with Mixup-based Distance Learning

Chaerin Kong  Jeesoo Kim  Donghoon Han  Nojun Kwak
Seoul National University
{veztylord, kimjiss0305, dhk1349, nojunk}@snu.ac.kr

Abstract

Producing diverse and realistic images with generative models such as GANs typically requires large scale training with vast amount of images. GANs trained with extremely limited data can easily overfit to few training samples and display undesirable properties like "stairlike" latent space where transitions in latent space suffer from discontinuity, occasionally yielding abrupt changes in outputs. In this work, we consider the situation where neither large scale dataset of our interest nor transferable source dataset is available, and seek to train existing generative models with minimal overfitting and mode collapse. We propose latent mixup-based distance regularization on the feature space of both a generator and the counterpart discriminator that encourages the two players to reason not only about the scarce observed data points but the relative distances in the feature space they reside. Qualitative and quantitative evaluation on diverse datasets demonstrates that our method is generally applicable to existing models to enhance both fidelity and diversity under the constraint of limited data. Code will be made public.

1. Introduction

Remarkable features of Generative Adversarial Networks (GANs) such as impressive sample quality and smooth latent space interpolation have drawn enormous attention from the community, but what we have enjoyed with little gratitude claim their worth in a data-limited regime. As naive training of GANs with small datasets often fails both in terms of fidelity and diversity, many have proposed novel approaches specifically designed for few-shot image synthesis. Among the most successful are those adapting a pre-trained source generator to the target domain [21,26,28] and those seeking generalization to unseen categories through feature fusion [13,16]. Despite their impressive synthesis quality, these approaches are often critically constrained in practice as they all require semantically related large source domain datasets to pretrain on. For some domains like abstract art paintings, medical images of rare symptoms and cartoon illustrations, it is very difficult or impossible to collect thousands of samples, while at the same time, finding an adequate source domain to transfer from is not straightforward either. This poses intimidating challenge for generative modeling, but we seek to find solutions in the darkest of times.

One of the biggest challenges of learning generative models with scarce data is that the model easily overfits. To directly address this issue, [46] and [18] have proposed data augmentation techniques that show promising results on low-shot generation tasks with datasets containing hundreds to thousands of training samples. Nevertheless, they show unsatisfactory performance with handful of data points (e.g., 10) and generative modeling under these circumstances still remains extremely challenging.

Inspired by [28] that proposes novel distance regularization to effectively transfer diversity information from the source to target, we recast the overfitting issue as stairlike latent space problem and suggest Mixup-Based Distance Learning (MDL) for both the generator and the discriminator to effectively control it.

As diversity is already heavily constrained by the small dataset, we wish to maximally exploit the given data points by continuously exploring their semantic mixups [44]. With overfitting, however, the discriminator is convinced only with the few observed samples, showing overly confident and abrupt decision boundaries. This induces the generator...
to map a subset of its latent space to a single mode, displaying stairlike transitions in the latent space. Hence, to *smooth* the feature space, we intentionally explore the generator latent space with continuous interpolation coefficient $c$, enforcing relative semantic distances between samples to follow the mixup ratio. Simultaneously, by penalizing the discriminator, we prohibit the discriminator from embedding images to arbitrary locations for its convenience of memorizing, and guide its feature space to be aligned with semantic distances.

We further observe that models trained with our regularizations resist mode collapse surprisingly well with neither large source domain datasets nor special augmentation. We believe that our distance regularizations encourage the model to preserve inherent diversity present in early stages throughout the course of training, opening up the doors for sample diversity under rigorous constraints.

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.

2. Related Works

**One-shot image generation** The most challenging case of few-shot image generation is the one-shot case where given a single image, the generator must learn its context to create diverse outcomes. SinGAN [32] approximates the distribution of a single image by differencing its scale during the adversarial training. Training from the downscaled image introduces ambiguity to the generator and this induces diversity when the generator comes to learn in high resolution. Based on SinGAN, ConSinGAN [15] proposes a technique to control the trade-off between fidelity and diversity of generated samples. One-Shot GAN [34] uses a dual-branch discriminator where each head respectively identifies real context and real layout of the generated sample. 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.

**Low-shot image generation** Given a limited amount of training data, the discriminator in conventional GAN can easily overfit. To mitigate this problem, DiffAugment [46] imposes differentiable data augmentation to both real and fake samples while ADA [18] devises non-leaking adaptive discriminator augmentation. As generating high resolution images is more challenging with limited data and budget, FastGAN [22] suggests a skip-layer excitation module and a self-supervised discriminator, which saves computational cost and stabilizes low-shot training. GenCo [9] 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.

**Transfer-based few-shot generation** Thus far, the few-shot image generation task where the dataset size is even more limited ($n \approx 10$) mostly required training on larger dataset with similar semantics [39, 40] mainly due to its inherent difficulty. A group of works [13, 16] learns transferable generation ability on seen categories and seek generalization into unseen categories through fusion-based methods. FreezeD [26] and EWC [21] introduce transfer learning frameworks for GANs that selectively update the target model’s parameters to mitigate overfitting. Meanwhile, CDC [28] 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 leverage the formulation of this work and modify for the situation where no source domain is available.

**Generative diversity** Mode collapse has been a long standing obstacle in GAN training. [2, 25] introduce divergence metrics that are effective at stabilizing GAN training while [10, 11] tackle this problem by training multiple networks. Another group of works [23, 24, 35, 42] proposes regularization methods to preserve distances in the generated output space. Unlike these works, we consider the few-shot setting where the phenomenon has a far devastating impact, and introduce an interpolation-based distance regularization method as an effective remedy.

**Latent mixup** Since [44], mixup-based methods have been actively explored in semi-supervised learning literature to enforce smooth behaviors in between training samples [3, 4, 36]. In generative models, [30] 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 [6, 20], it has been mainly studied with autoencoders. [5, 29] regularize autoencoders to learn smooth latent space while [31, 41] explore their potential as generative models through interpolation.

3. Approach

We consider the situation where only few train examples (e.g., $n = 10$) are available with no semantically sim-
Figure 2. Overview of our Mixup-based Distance Learning (MDL). We sample interpolation coefficients from a Dirichlet distribution and generate an anchor point $z_0$ through interpolation. Then we enforce pairwise similarities between intermediate generator activations to follow the interpolation coefficients. Similar regularization is imposed on discriminator’s penultimate activation, which is linearly projected before similarity calculation. The proposed regularization terms can be added on top of any traditional adversarial framework.

ilar source domain. Under harsh data constraint, overfitting greatly restricts a model’s ability to learn data distribution and produce diverse samples. We identify its byproduct un-smooth 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 (MDL) 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 [28], 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 [28], 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

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

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

where $G^l$ is the generator activation at the $l$th layer and $\{z_i\}_i^N$ are latent vectors. Note that $G_s$ and $G_{s\rightarrow t}$ 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 $\mathcal{L}_{adv}$, they impose a KL-divergence-based regularization of the following form:

$$
\mathcal{L}_{dist} = E_{z \sim p_s(z)}[D_{KL}(q^l || p^l)].
$$

The benefits of this auxiliary objective are twofold: it prevents distance collapse in the target domain through distribution binding and transfers diversity from the source to target via one-to-one correspondence.

3.2. Generator Latent Mixup

In [28], the anchor point $z_0$ could be chosen arbitrarily from the prior distribution $p_s(z)$ 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 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:

$$
z_0 = \sum_{i=1}^{N} c_i z_i, \quad c \sim \text{Dir}(\alpha_1, \ldots , \alpha_N)
$$

where $c \triangleq [c_1, \ldots , c_N]^T$. Using Eq. (4), the latent space can be navigated in a quantitatively controlled manner. Defining probability distribution of pairwise similarities as
in [28], we bind it to the interpolation coefficients \( c \) instead. The proposed distance loss is defined as follows:

\[
\mathcal{L}_{dist}^G = \mathbb{E}_{z \sim p_z(z), c \sim \text{Dir}(1)}[D_{KL}(q || p)],
\]

\[
q_i = \text{softmax}(\{\sin(G^i(z_0), G^i(z_1))\}_{i=1}^N),
\]

\[
p = \text{softmax}(\{c_i\}_{i=1}^N),
\]

where \( \text{Dir}(1) \) denotes the Dirichlet distribution with all-1 parameters. This efficiently accomplishes our two desiderata. Intuitively, unlike naive generators that gradually converge to a 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 converge to few modes, our regularization forces the generated data. Intuitively, unlike naive generators that gradually converge to a 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 [28], explicitly enforcing smooth latent interpolation. An anchor point similar to [28] can be obtained with one-hot coefficients \( c \).

### 3.3. Discriminator Feature Space Alignment

While the generator distance regularization alone 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 distance 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 \( G(z_i) \) is provided to \( D \), we construct an \( N \)-way distribution similar to Eq. (6) as

\[
r = \text{softmax}(\{\text{proj}(d_1(z_i)), \text{proj}(d_1(z_1))\}_{i=1}^N)
\]

where \( \text{proj} \) refers to a linear projection layer widely used in self-supervised learning literature [7, 8, 12] and \( d_1(z_0) \). 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

\[
\mathcal{L}_{dist}^D = \mathbb{E}_{z \sim p_z(z), c \sim \text{Dir}(1)}[D_{KL}(r || p)].
\]

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 discriminator to provide smoother and more semantically meaningful signals to the generator.

### 3.4. Final Objective

Fig. 2 shows an overall concept of our method. Our final objective for the generator and the discriminator takes the form:

\[
\mathcal{L}^G = \mathcal{L}_{adv}^G + \lambda_G \mathcal{L}_{dist}^G
\]

\[
\mathcal{L}^D = \mathcal{L}_{adv}^D + \lambda_D \mathcal{L}_{dist}^D
\]

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, StyleGAN2 \(^1\) [20] and FastGAN [22]. We keep their objective functions as they are and simply add our regularization terms. For StyleGAN2, we interpolate in \( \mathcal{V} \) rather than \( \mathcal{Z} \), which has been shown to have better properties such as disentanglement [1, 37, 47]. Interpolation coefficients \( c \) is sampled from a Dirichlet distribution of parameters all equal to one. Patch-level discrimination [17, 28] is applied for interpolated images to encourage our generator to be creative while exploring the latent space.

### 4. Experiments

We validate our method on diverse domains, ranging from photorealistic images to hand-drawn illustrations. We further analyze the effect of individual components, i.e., model architecture, data augmentation and proposed regularizations, through ablation on different datasets, demonstrating that our approach is compatible with existing model architectures and augmentation techniques.

**Baselines** We mainly apply our method to the state-of-the-art unconditional GAN model, StyleGAN2 [20]. Data augmentation techniques introduced by [46] and [18] show promising performance on low-shot image generation task, so we evaluate them along with ours and refer to them as DiffAug and ADA respectively. As we observed significant performance drops for DiffAug when applied to normal StyleGAN2 with 8-layer mapping network, we set the number of FC layers to 2 only for DiffAug as done by the authors while leaving the base architecture untouched for others. We additionally apply our method to FastGAN [22], which is a light-weight GAN architecture based on novel Skip-Layer channelwise Excitation module (SLE-module) that allows faster convergence with limited data.

**Datasets** For quantitative evaluation, we use Animal-Face Dog [33], Oxford-flowers [27], FFHQ-babies [19], face sketches [38], 100-shot Obama and Grumpy Cat [46], anime face [22] and Pokemon (pokemon.com, [22]). 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 face paintings of Amedeo Modigliani [43] and landscape drawings [28]. All experiments are done on

---

\(^1\)https://github.com/rosinality/stylegan2-pytorch
Figure 3. 10-shot image generation results. Baseline methods either collapse to few modes or simply replicate the training samples. Note that we can find corresponding real sample for each well-synthesized image for the baselines. Our method actively encourages the generator to explore different semantic mixups of given samples, which enables synthesis of various unseen samples.

256 × 256 images. Tab. 1 summarizes the datasets and the number of shots used in each dataset. Generation results on other datasets can be found in the supplementary materials.

**Evaluation Metrics** One of the difficulties in few-shot generative modeling lies in the subtlety of evaluation, as the most widely used evaluation metric, Fréchet Inception Distance (FID) [14], does not accurately represent generation quality [21] with small datasets. We measure FID for datasets containing a sufficient number (≥ 100) of samples along with pairwise Learned Perceptual Image Patch Similarity (LPIPS) [45] as a measure of diversity. For simulated few-shot tasks, FID score is computed against the full dataset, as in [21, 28]. We further use LPIPS as a distance metric for demonstrating interpolation smoothness and mode preservation.

Table 1. Number of shots used in each dataset. **Names of datasets** are presented in the first and third rows and their corresponding **number of shots** used in this paper are described in the second and fourth rows.
4.1. Qualitative Result

Fig. 3 shows generated samples of different methods trained with 10 real samples. We observe that baseline methods either collapse to few modes or severely overfit to the training data, resulting in inability to generate diverse novel samples. Ours is the only method that produces a variety of convincing samples that are not present in the training set. We can see that 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. 4 we display “uncurated” sets of images generated from models trained with 10-shot paintings of Amedeo Modigliani that share a common training sample as nearest neighbor. Outputs of DiffAug show little differences among them, 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. More generated samples from Amedeo Modigliani paintings can be found in the supplementary materials.

4.2. Quantitative Evaluation

Tab. 2 shows FID and LPIPS scores for 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.

While pretraining-free 10-shot image synthesis task has not been studied much, several works [22, 46] have previously explored generative modeling with as little as 100 samples. We present quantitative evaluations on popular low-shot benchmarks in Tab. 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.

4.3. Ablation Study

We further evaluate the effects of the two proposed regularizations, MDL-G (for generator) and MDL-D (for discriminator), through ablation under different settings. In Tab. 4, we observe that in general, our regularizations both contribute to better image quality and diversity, while in some special cases, only adding MDL-G leads to better FID score. We conjecture that aligning discriminator’s penultimate output vectors with the interpolation coefficients can impose overly strict constraint for some datasets. We nonetheless observe consistent improvements on diversity.

Tab. 5 shows the performance of our method under a larger data setting. Since FFHQ-babies contains more than 2,000 images, we randomly sample subsets of size 10,
Table 4. Ablation experiments for MDL-G and MDL-D. We observe that the two regularizations combined generally yields the best performances.

| Base Method | MDL | Dog (10-shot) | Babies (100-shot) | Flowers (100-shot) |
|-------------|-----|---------------|-------------------|-------------------|
| StyleGAN2   |     | 177.94 0.5688 | 130.99 0.5744     | 192.15 0.7472     |
| ✓           |     | 95.36 0.6731  | 71.70 0.6381      | 84.03 0.7801      |
| ✓ ✓         |     | 97.82 0.6412  | 63.55 0.6646      | 81.98 0.7823      |

Table 5. Shot ablation results on FFHQ-babies. Ours consistently improves the baseline in the data-limited regime.

| Method       | Metric  | k-shot | 10 | 100 | 1000 |
|--------------|---------|--------|----|-----|------|
| StyleGAN2    | FID (↓) | 184.77 | 130.99 | 44.42 |
|              | LPIPS (↑)| 0.4756 | 0.5744 | 0.6300 |
| StyleGAN2 + Ours | FID (↓) | 93.15  | 63.35  | 38.05  |
|              | LPIPS (↑)| 0.6230 | 0.6468 | 0.6685 |

Table 6. Ablation for Dirichlet parameters on Anime Face.

| Metric | α = 0.1 | α = 1  | α = 10 |
|--------|---------|--------|--------|
| FID (↓)| 76.35   | 73.16  | 80.83  |
| LPIPS (↑)| 0.5359 | 0.5484 | 0.5321 |

Table 7. Perceptual Path Length uniformity. We subdivide a latent interpolation path into 10 subintervals and compute perceptual distances. Standard deviation is computed across the subintervals, indicating perceptual uniformity of latent transition.

| Method       | Mean   | Std. Dev. | Endpoint Mean |
|--------------|--------|-----------|---------------|
| StyleGAN2    | 21.910 | 12.657    | 60.901        |
| StyleGAN2 + ADA | 25.577 | 9.198     | 63.558        |
| StyleGAN2 + DiffAug | 0.820 | 0.095     | 22.083        |
| StyleGAN2 + Ours | 12.818 | 4.187     | 64.280        |

We can see that the performance of StyleGAN2 steadily improves with more training samples, but our method can consistently push the curve upward. 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 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 [19].

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 [19]. 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 and standard deviation, and the mean for the full interval. Note that as PPL is a quadratic measure, the sum of subinterval means can be smaller than the endpoint mean. We first notice that StyleGAN2+DiffAug suffers severe mode collapse as apparent from Fig. 5, so the low mean and standard deviation values simply indicate training failures. StyleGAN2 with no augmentation, StyleGAN2+ADA and ours show similar endpoint mean, suggesting that the overall total perceptual distance is consistent, while ours displays the lowest mean PPL and 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. 5 where the baselines either collapse or display stairlike latent transition while ours shows smooth semantic interpolation. More details on PPL computation can be found in the supplementary materials.

4.5. Preserving Diversity

As opposed to [28] 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/difference 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. 6 shows the result, where the number of modes is represented by the number of unique training samples (real images) that are the nearest neighbor
Figure 6. Analysis on sample diversity. (a) shows that our method produces samples with greater diversity. (b) indicates the number of unique training samples that are nearest neighbor to any of the generated samples. We generate 500 samples for the analysis. Since we train our model with 10 samples from anime face dataset, the upper bound is 10. Training snapshots are available in the supplementary materials.

to any of the generated images. In Fig. 6a, we can see that vanilla StyleGAN2 and our method show similar LPIPS in the beginning, but the baseline quickly loses diversity as opposed to ours that maintain relatively high level of diversity throughout the training. Fig. 6b delivers similar implication that FastGAN trained with our method better preserves modes, thus diversity, compared to the baseline.

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

5. Discussion

The trade-off between fidelity and diversity in GANs has been noted by many [6, 19]. 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 Oxford-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, our proposed method shows significant contributions, confirming that it excels at learning diversity even under challenging situations.

6. Conclusion

Image generation with a handful of data samples has been deemed extremely challenging due to overfitting and mode collapse, thus it was mainly tackled with transfer learning based methods. However, these approaches share an inherent limitation that a semantically similar source domain dataset should be available. To overcome this limitation and broaden the potential applications for GAN models, we propose mixup-based distance regularizations that effectively alleviate mode collapse and smooth the otherwise stairlike latent space of generative models. We empirically demonstrate that our method is capable of generating diverse unseen samples of convincing quality after training on as little as 10 real samples, with no pretraining or special data augmentation. In short, our method can be directly added on top of an existing GAN model, e.g., StyleGAN2, to accommodate it for few-shot image synthesis. We hope our work facilitates future research on data efficient generative modeling, which we believe has great upside in both academics and applications.
References

[1] Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. Restyle: A residual-based stylegan encoder via iterative refinement. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 6711–6720, 2021. 4

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In International conference on machine learning, pages 214–223. PMLR, 2017. 2

[3] David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution alignment and augmentationanchoring. arXiv preprint arXiv:1911.09785, 2019. 2

[4] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. arXiv preprint arXiv:1905.02249, 2019. 2

[5] David Berthelot, Colin Raffel, Aurko Roy, and Ian Goodfellow. Understanding and improving interpolation in autencoders via an adversarial regularizer. arXiv preprint arXiv:1807.07543, 2018. 2

[6] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. arXiv preprint arXiv:1809.10196, 2018. 2, 8

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In International conference on machine learning, pages 1597–1607. PMLR, 2020. 4

[8] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 15750–15758, 2021. 4

[9] Kaizhen Cui, Jiexing Huang, Zhipeng Luo, Gongjie Zhang, Fangneng Zhan, and Shijian Lu. Geno: Generative co-training on data-limited image generation. arXiv preprint arXiv:2110.01254, 2021. 2

[10] Ishan Durugkar, Ian Gemp, and Sridhar Mahadevan. Generative multi-adversarial networks. arXiv preprint arXiv:1611.01673, 2016. 2

[11] Arnab Ghosh, Viveka Kulharia, Vinay P Namboodiri, Philip HS Torr, and Puneet K Dokania. Multi-agent diverse generative adversarial networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 8513–8521, 2018. 2

[12] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. arXiv preprint arXiv:2006.07733, 2020. 4

[13] Zheng Gu, Wenbin Li, Jing Huo, Lei Wang, and Yang Gao. Lofigan: Fusing local representations for few-shot image generation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 8463–8471, 2021. 1, 2

[14] 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. Advances in neural information processing systems, 30, 2017. 5

[15] Tobias Hinz, Matthew Fisher, Oliver Wang, and Stefan Wermter. Improved techniques for training single-image gans. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 1300–1309, 2021. 2

[16] Yan Hong, Li Niu, Jianfu Zhang, and Liqing Zhang. Matchinggan: Matching-based few-shot image generation. In 2020 IEEE International Conference on Multimedia and Expo (ICME), pages 1–6. IEEE, 2020. 1, 2

[17] Philipp Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1125–1134, 2017. 4

[18] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. arXiv preprint arXiv:2006.06676, 2020. 1, 2, 4

[19] 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. 4, 7, 8

[20] 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. 2, 4

[21] Yijun Li, Richard Zhang, Jingwen Lu, and Eli Shechtman. Few-shot image generation with elastic weight consolidation. arXiv preprint arXiv:2012.02780, 2020. 1, 2, 5

[22] Bingchen Liu, Yizhe Zhu, Kunpeng Song, and Ahmed Elgammal. Towards faster and stabilized gan training for high-fidelity few-shot image synthesis. In International Conference on Learning Representations, 2020. 2, 4, 6

[23] Shaohui Liu, Xiao Zhang, Jianqiao Wangni, and Jianbo Shi. Normalized diversification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10306–10315, 2019. 2

[24] Qi Mao, Hsin-Ying Lee, Hung-Yu Tseng, Siwei Ma, and Ming-Hsuan Yang. Mode seeking generative adversarial networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1429–1437, 2019. 2

[25] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhe Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1300–1309, 2020. 2

[26] Sangwoo Mo, Minsu Cho, and Jinwoo Shin. Freeze the discriminator: a simple baseline for fine-tuning gans. arXiv preprint arXiv:2002.10964, 2020. 1, 2

[27] M-E Nilsback and Andrew Zisserman. A visual vocabulary for flower classification. In 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’06), volume 2, pages 1447–1454. IEEE, 2006. 4
[28] Utkarsh Ojha, Yijun Li, Jingwan Lu, Alexei A Efros, Yong Jae Lee, Eli Shechtman, and Richard Zhang. Few-shot image generation via cross-domain correspondence. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10743–10752, 2021. 1, 2, 3, 4, 5, 7

[29] Alon Oring, Zohar Yakhini, and Yacov Hel-Or. Autoencoder image interpolation by shaping the latent space. arXiv preprint arXiv:2008.01487, 2020. 2

[30] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434, 2015. 2

[31] Tim Sainburg, Marvin Thielk, Brad Theilman, Benjamin Migliori, and Timothy Gentner. Generative adversarial interpolative autoencoding: adversarial training on latent space interpolations encourage convex latent distributions. arXiv preprint arXiv:1807.06650, 2018. 2

[32] Tamar Rott Shaham, Tali Dekel, and Tomer Michaeli. Simgan: Learning a generative model from a single natural image. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 4570–4580, 2019. 2

[33] Zhangzhang Si and Song-Chun Zhu. Learning hybrid image templates (hit) by information projection. IEEE Transactions on pattern analysis and machine intelligence, 34(7):1354–1367, 2011. 4

[34] Ngoc-Trung Tran, Tuan-Anh Bui, and Ngai-Man Cheung. Dist-gan: An improved gan using distance constraints. In Proceedings of the European Conference on Computer Vision (ECCV), pages 370–385, 2018. 2

[35] Vikas Verma, Kenji Kawaguchi, Alex Lamb, Juho Kannala, Arno Solin, Yoshua Bengio, and David Lopez-Paz. Interpolation consistency training for semi-supervised learning. Neural Networks, 2021. 2

[36] Peihao Zhu, Rameen Abdal, Yipeng Qin, John Femiani, and Peter Wonka. Improved stylegan embedding: Where are the good latents? arXiv preprint arXiv:2012.09036, 2020. 4

[37] Xiaogang Wang and Xiaoou Tang. Face photo-sketch synthesis and recognition. IEEE transactions on pattern analysis and machine intelligence, 31(11):1955–1967, 2008. 4

[38] Yanxi Wang, Abel Gonzalez-Garcia, David Berga, Luis Herranz, Fahad Shabbaz Khan, and Joost van de Weijer. 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, pages 9332–9341, 2020. 2

[39] Yong Zhang, Seunghoon Hong, Yunsook Jang, Tianchen Zhao, and Honglak Lee. Diversity-sensitive conditional generative adversarial networks. arXiv preprint arXiv:1901.09024, 2019. 2
A. Implementation Details

**StyleGAN2** We adopt the standard StyleGAN2 architecture for 256 × 256 resolution images, with 8 fully connected layers in the mapping network. We keep the hyper-parameters such as the learning rate, regularization weights and frequency, untouched, and only add our proposed MDL.

**DiffAug** We essentially follow the official configuration for low-shot generation, including the two-layer mapping network and three data augmentation methods. We have also tried with a standard 8 FC layer mapping network and observed significant drops in the overall performance as shown in Tab. S1.

| FC layers | Obama (100-shot) | Grumpy Cat (100-shot) |
|-----------|------------------|-----------------------|
| 2         | 46.87            | 26.52                 |
| 8         | 71.13            | 38.42                 |

**FastGAN** We use the official FastGAN implementation for 256 × 256 images. As FastGAN doesn’t have a separate mapping network, we interpolate in Z space.

**MDL** For MDL, we alternate between the normal adversarial training step and the interpolation/regularization step. In the former we go through normal image-level discrimination and in the latter, we apply patch-level discrimination on the mixup samples and compute losses for MDL-G and MDL-D. For patch discrimination, we largely adopt the implementation of Cross-domain Correspondence (CDC). Our linear projection layer for the discriminator operates on 512 dimension.

**Perceptual Path Length** For PPL computation, we mainly follow the implementation in StyleGAN. The difference is that we subdivide a latent interpolation path into 10 subintervals and compute the perceptual distance for each line segment. Since the original PPL computation divides the perceptual distance by the squared step size, we divide each subinterval length by 0.1². For clear demonstration, we divide the endpoint mean by 0.1² as well. Note that the overall procedure is equivalent to calculating LPIPS multiplied by the factor of 100. The standard deviation is computed across the subintervals, and averaged for the interpolation paths.

**Number of Modes** We generate 500 samples and compute their perceptual distances to the 10 training samples. We record the index for the real sample with the smallest perceptual distance and report the unique count. It is visually apparent from Fig. S1 that our method helps the model preserve modes and maintain diversity.

---

B. Training Snapshots

We provide training snapshots for FastGAN and StyleGAN2 for visual demonstration of diversity and interpolation smoothness. Fig. S1 clearly shows that as opposed to vanilla FastGAN that rapidly loses diversity and converges to few prototypes, MDL successfully alleviates this. Fig. S2 displays interpolation snapshots for StyleGAN2. In early training iterations, it does show relatively smooth latent transition, but the sample quality is very unsatisfactory. As the training proceeds, the sample quality improves as the model overfits, but consequently the interpolation smoothness is quickly lost. This describes the classic dilemma in few-shot generative modeling. In contrast, Fig. S3 shows that as MDL is effective at maintaining latent space smoothness, it provides a sweet spot where reasonable sample quality and smooth latent transition coexist. Note that models with MDL do inevitably overfit in the end, but we can find reasonable stopping point that produces diverse unseen samples with satisfactory visual quality.

C. Additional Generated Samples

We present synthesis results from face paintings of Amedeo Modigliani, abstract paintings of Paul Klee and illustrations of Japanese animation character Totoro in Fig. S4, Fig. S5 and Fig. S6, respectively. Images of Paul Klee and Totoro were crawled from the web, and we used only 5 real samples for the latter. Additional interpolation examples are also displayed in Fig. S7 and Fig. S8.

D. Sample Images from Low-shot Benchmarks

In Fig. S9, we present samples from Obama and Grumpy Cat datasets. As they contain images of a single character, the intra-diversity is inherently very limited.

E. Naive Application of GAN adaptation

We display results from naive application of CDC. Since it is very difficult to find a semantically similar source domain for datasets like Pokemon and abstract paintings of Paul Klee, we naively leverage the source generator trained on FFHQ. As the source and the target are semantically different, the adaptation does not yield satisfactory outcomes as expected. We can observe the dilemma here as well that in the early iterations, the face shape learned in the source domain is clearly visible while in later stages, the face shape is no longer visible but the model collapses altogether. As CDC preserves distances in the target domain through the correspondence to the source domain, it is not applicable to domains that lack an adequate source dataset to transfer from.

---

2 https://github.com/rosinality/stylegan2-pytorch
3 https://github.com/mit-han-lab/data-efficient-gans
4 https://github.com/odegeasslbc/FastGAN-pytorch
5 https://github.com/utkarshoja/few-shot-gan-adaptation
6 https://www.kaggle.com/robgonsalves/abstract-paintings/version/1, CC BY 4.0
Figure S1. Training snapshots for FastGAN and FastGAN+MDL in early iterations. As opposed to the base FastGAN that rapidly loses diversity, our regularizations help preserve the modes throughout the course of training. Numbers in the left indicate training iterations.
Figure S2. Interpolation snapshots for StyleGAN2. Numbers in the left indicate training iterations.
Figure S3. Interpolation snapshots for StyleGAN2+MDL.

Figure S4. Samples from face paintings of Amedeo Modigliani. While the baselines simply replicate the given images, ours produces diverse unseen face images. *Ours* represents samples from StyleGAN2+MDL.
Figure S5. Samples from abstract paintings of Paul Klee. With abstract paintings, it is particularly not straightforward to find an adequate source domain for transfer learning. Generated represents images from StyleGAN2+MDL.

Figure S6. Result from 5-shot training on Totoro. Although there are only 5 training samples, ours combines visual features in a natural way to produce diverse novel samples.

Figure S7. More interpolation examples. Numbers in the parentheses represent the number of training samples used for each dataset.
Figure S8. Interpolation examples. Baselines clearly display *stairlike* latent transition while ours shows smooth interpolation.

Figure S9. Random samples from low-shot benchmark datasets, Obama and Grumpy Cat. Since they contain photos of a single character, the intra-diversity is inherently constrained.
Figure S10. Naive application of CDC from FFHQ to Pokemon. As the authors have pointed out, the adaptation performance degrades when the two domains are semantically different, but it is not straightforward to find a transferable source domain for datasets like Pokemon. We observe clear human face shapes in the early stages (left) and mode collapse in later stages (right) where the face shape is no longer visible.

Figure S11. Naive application of CDC from FFHQ to paintings of Paul Klee. As in Fig. S10, the model totally collapses as it discards the face shape prior learned in the source domain. Because CDC relies on correspondence between the source and the target, it is not effective when a semantically similar source domain dataset is not available.