MS^{3}D: A RG Flow-Based Regularization for GAN Training with Limited Data

Jian Wang [1] **Xin Lan** [1] **Yuxin Tian** [1] **Jiancheng Lv** [1]

Abstract

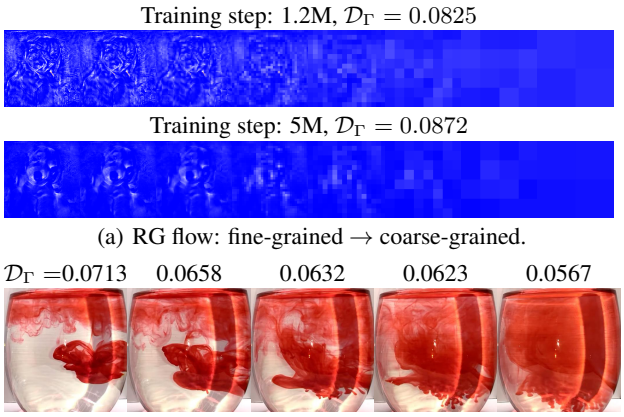
Generative adversarial networks (GANs) have made impressive advances in image generation, but they often require large-scale training data to avoid degradation caused by discriminator overfitting. To tackle this issue, we investigate the challenge of training GANs with limited data, and propose a novel regularization method based on the idea of renormalization group (RG) in physics.We observe that in the limited data setting, the gradient pattern that the generator obtains from the discriminator becomes more aggregated over time. In RG context, this aggregated pattern exhibits a high discrepancy from its coarse-grained versions, which implies a high-capacity and sensitive system, prone to overfitting and collapse. To address this problem, we introduce a **m**ulti-**s**cale **s**tructural **s**elf-**d**issimilarity (MS^{3}D) regularization, which constrains the gradient field to have a consistent pattern across different scales, thereby fostering a more redundant and robust system. We show that our method can effectively enhance the performance and stability of GANs under limited data scenarios, and even allow them to generate high-quality images with very few data.

1. Introduction

The challenge of training GANs with limited data has garnered increasing attention within the research community. Numerous studies (Zhao et al., 2020a; Karras et al., 2020a; Jiang et al., 2021; Tseng et al., 2021; Fang et al., 2022; Cui et al., 2023) have reached a consensus: insufficient data often leads to overfitting in the discriminator. This overfitting results in a lack of meaningful dynamic guidance for the generator, ultimately causing its performance to degrade.

[1]College of Computer Science, Sichuan University and Engineering Research Center of Machine Learning and Industry Intelligence, Ministry of Education, Chengdu 610065, P.R. China. Correspondence to: Jiancheng Lv <lvjiancheng@scu.edu.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).



(a) RG flow: fine-grained → coarse-grained.

(b) Snapshots of food coloring diffusion in hot water over time.

Figure 1. Illustrative examples of renormalization group (RG) flow. The aggregated pattern exhibits high multi-scale structural self-dissimilarity (MS^{3}D), denoted by $\mathcal{D}_\Gamma$.

Data augmentation, a standard solution to prevent overfitting in deep learning (Shorten & Khoshgoftaar, 2019), has been widely applied to GAN training in limited data scenarios. Traditional (Zhao et al., 2020b; Tran et al., 2021) and differentiable (Zhao et al., 2020a) data augmentation techniques for both real and generated images, as well as various adaptive augmentation strategies (Karras et al., 2020a), have shown good performance on several standard benchmarks. However, the effectiveness of data augmentation heavily depends on specific handcrafted types or a costly search process, which limits its generality (Zhao et al., 2020b; Karras et al., 2020a). Additionally, data augmentation addresses data deficiency by enhancing the quantity and diversity of samples without providing a deeper understanding of the internal dynamics of GANs.

In this paper, we address the problem of GAN deterioration from a novel perspective and introduce a model regularization method. Unlike most existing regularization techniques (Tseng et al., 2021; Cui et al., 2022; 2023) that focus on improving the discriminator's generalization across different data distributions to avoid overfitting, our approach explores the intrinsic properties of neural networks to uncover potential clues. Specifically, we observe that in settings with limited data, the gradients provided by the discriminator, i.e., $\nabla_x f(x;\phi)$, gradually exhibit an aggregation pattern. This pattern indicates that the discriminator concentrates its attention on a small portion of the input

image, rather than capturing comprehensive details. This phenomenon, which we term the *perceptual narrowing phenomenon*, varies in degree across different GANs and tasks and accompanies overfitting.

To understand this phenomenon, here, we draw inspiration from the renormalization group (RG) concept in physics (Kadanoff, 1966; Wilson, 1971), which progressively separates coarse-grained statistics from fine-grained statistics through local transformations at different scales. We apply this idea to analyze the gradient field $\nabla_x f(x;\phi)$ at different scales. Unlike scatter patterns, the aggregated pattern shows significant divergence from its coarse-grained versions, as depicted in Fig. 1. This *self-dissimilarity* (SD)[1] can be seen as the system's unique "signature" (Wolpert & Macready, 2018; 2007), revealing how information processing changes across different scales in the system (Jacobs & Jacobs, 1992; Wolpert & Macready, 2007). A high SD indicates that the system is efficient, encoding substantial processing into its dynamics, but also sensitive or unstable, where a small change can cause large dynamic shifts. Conversely, a low SD is associated with robustness, where a system can maintain functionality despite disturbances, being less efficient but more reliable due to built-in redundancies. The latter is more desirable for GAN training.

Based on this analysis, we introduce a **m**ulti-**s**cale **s**tructural **s**elf-**d**issimilarity (MS^{3}D) regularization based on RG flow. By repeatedly applying RG transformations, we generate a series of coarse-grained versions at different scales. Along the RG flow, we compute SD at each scale and combine them to obtain MS^{3}D. This regularization enforces the gradient field to maintain a similar pattern or structure across scales, promoting a more redundant and robust feedback system for the generator. These properties help avoid overfitting and collapse in GAN training with limited data. Crucially, the proposed MS^{3}D computation method is differentiable and can be readily implemented using popular deep learning frameworks like PyTorch (Paszke et al., 2019) and TensorFlow (Abadi et al., 2016). We verify the effectiveness of our method on various GANs and datasets. The results demonstrate that it improves GAN performance under limited data conditions in terms of generalization and stability. Notably, our method is orthogonal to data augmentation methods and other model constraint techniques, and can be integrated with them to further enhance GAN performance with very small datasets. Our work provides a new perspective on GAN training with limited data and offers a novel regularization method to improve GAN performance in this challenging scenario.

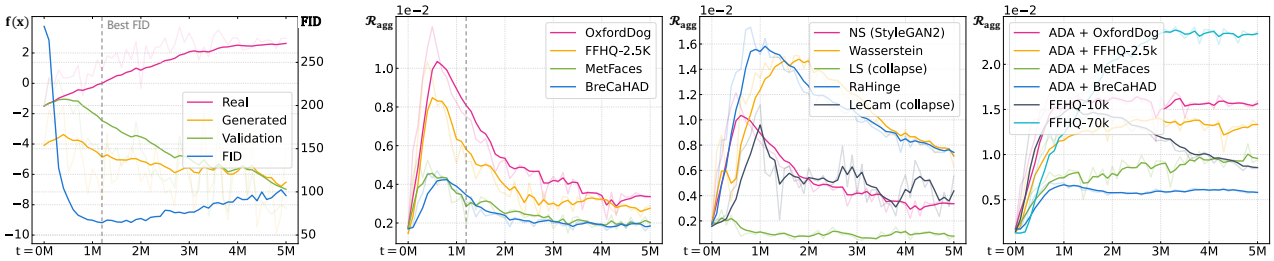
[1] A concept in complexity theory and the first formalization of SD appeared in (Wolpert & Macready, 2018).

2. Related Work

Generative adversarial networks. Generative adversarial networks (GANs) (Goodfellow et al., 2014) have made significant advancements in generating high-quality and diverse images. This progress is attributed to the development of more robust objective functions (Arjovsky et al., 2017; Gulrajani et al., 2017; Zhao et al., 2017; Mao et al., 2017; Song & Ermon, 2020), advanced architectures (Miyato et al., 2018; Miyato & Koyama, 2018; Zhang et al., 2019), and effective training strategies (Denton et al., 2015; Zhang et al., 2017; Karras et al., 2018; Liu et al., 2020). Notable examples include BigGAN (Brock et al., 2019) and StyleGAN (Karras et al., 2019; 2020b), which are capable of generating high-resolution images with rich details and diverse styles. This paper focuses on the challenges and solutions for training GANs with limited data.

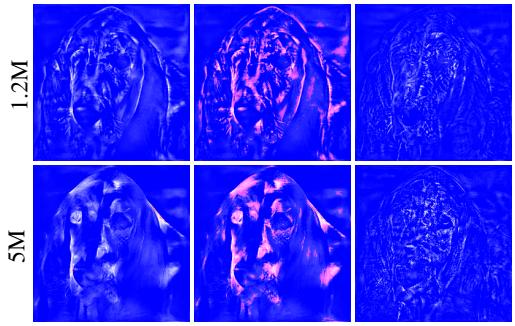
Training GANs under limited data setting. Training GANs with limited data presents significant challenges, as the discriminator may overfit, leading to degraded generated samples (Webster et al., 2019; Gulrajani et al., 2019). One common approach to mitigate this issue is data augmentation (Karras et al., 2020a; Tran et al., 2020; Zhang et al., 2020; Zhao et al., 2020a; 2021; 2020b), which enriches the data distribution by applying transformations to the original samples. However, data augmentation is not straightforward for GANs, as it can alter the target distribution or introduce artifacts (augmentation leaking). Recent methods have been designed to address these challenges, such as differentiable augmentation (Zhao et al., 2020a), adaptive augmentation (Karras et al., 2020a), and generative augmentation (Zhao et al., 2020b).

Another approach is model regularization, which prevents the discriminator from overfitting by imposing constraints or penalties on its parameters or outputs. While model regularization is commonly used to stabilize GAN training and prevent mode collapse, it is particularly effective in data-limited settings where overfitting is more severe. Techniques include adding noise to the discriminator's inputs or outputs (Arjovsky & Bottou, 2017; Sønderby et al., 2017; Jenni & Favaro, 2019), applying gradient penalties (Gulrajani et al., 2017; Mescheder et al., 2018), using spectral normalization (Miyato et al., 2018; Miyato & Koyama, 2018), and adding consistency loss (Zhang et al., 2020). Recent innovations, such as the LC regularization term (Tseng et al., 2021), which modulates the discriminator's evaluations using two exponential moving average variables and connects to LeCam divergence (Cam, 1986), have shown significant benefits. DigGAN (Fang et al., 2022) addresses gradient discrepancies between real and generated images, improving GAN performance. Additionally, leveraging external knowledge by using pre-trained models as additional discriminators (Kumari et al., 2022), aligning discriminator features



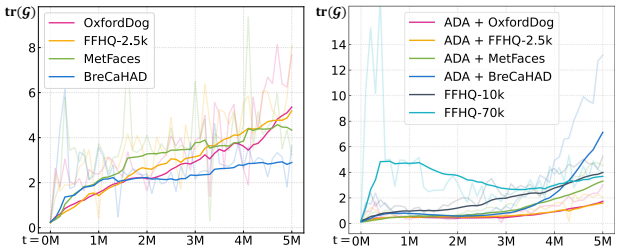
(a) Discriminator outputs, FIDs (b) The aggregation degree of $\nabla_x f(x;\phi)$

Figure 2. (a) Evolution of discriminator outputs and FID values during training, illustrating the discriminator's overfitting and the resulting degradation of generated samples. (b) Consistent aggregation tendency of the gradient $\nabla_x f(x;\phi)$ across various GANs and datasets under limited data conditions.



(a) StyleGAN2 (b) + ADA

Figure 3. Visualization of the gradient $\nabla_x f(x;\phi)$ at different training steps.



(a) (b)

Figure 4. (a) Under limited data settings, Fisher information increases during training, which indicates a decline in system stability. (b) When data augmentation is applied or the data volume is increased, Fisher information remains low, suggesting enhanced system stability.

through knowledge distillation (Cui et al., 2023), or pre-training the highest resolution layer on larger datasets (Mo et al., 2020) has also proven effective.

Our work introduces a novel model regularization technique, offering deeper insights into GANs' internal dynamics through the lens of the renormalization group.

Renormalization group and its applications. The renormalization group (RG) is a powerful tool in modern physics that elucidates how physical systems exhibit distinct behaviors at different observational scales (Kadanoff, 1966; Wilson, 1971). RG is extensively used in various physics fields, such as condensed matter theory (Wilson, 1971; Fisher, 1998; Stanley, 1999), quantum field theory (Machacek & Vaughn, 1985), and complexity theory (Bagrov et al., 2020). Moreover, the concept of RG has been linked to information theory (Koch-Janusz & Ringel, 2018; Gordon et al., 2021) and principal component analysis (Bradde & Bialek, 2017), making them more accessible to machine learning researchers. RG has also been employed to study neural networks' properties and dynamics. For example, Mehta & Schwab (2014) demonstrated that RG can map a restricted Boltzmann machine to a hierarchical model, and Lin et al. (2017) showed that RG can explain the good generalization performance of deep neural networks. In this paper, we introduce the first application of RG to GAN training, proposing a novel regularization technique inspired by RG principles.

3. Methodology

3.1. Generative Adversarial Networks

Generative adversarial networks (GANs) (Goodfellow et al., 2014) are a class of generative models designed to synthesize realistic data samples from a latent noise vector z. GANs consist of two neural networks: a generator $g(\cdot;\theta)$ that transforms the noise vector into a data sample, and a discriminator $f(\cdot;\phi)$ that distinguishes between real and generated samples. These networks are trained adversarially; the generator aims to produce samples that mimic the real data distribution, while the discriminator aims to accurately classify samples as real or fake. The training objective of GANs is formulated as a minimax game:

$$\min_{\theta}\max_{\phi}\mathbb{E}_x[\log f(x;\phi)]+\mathbb{E}_z[\log(1-f(g(z;\theta);\phi))]. \quad (1)$$

The optimal solution is a Nash equilibrium where the generator produces samples indistinguishable from real data, and the discriminator assigns a probability of 0.5 to all samples.

3.2. Perceptual Narrowing Phenomenon

Discriminator overfitting is a critical issue in GAN training with limited data, leading to a degradation in the quality of generated images (Zhao et al., 2020a; Karras et al., 2020a; Jiang et al., 2021; Tseng et al., 2021). This phenomenon is demonstrated in our experiments with StyleGAN2 (Karras et al., 2020b) on OxfordDog dataset (Parkhi et al., 2012), as shown in Fig. 2(a). The discriminator becomes increasingly confident about the real images from the training set while becoming less confident about real images from the validation set, leading to a deterioration in Fréchet Inception Distance (FID) (Heusel et al., 2017) over time.

Additionally, we observe that the gradient field $\nabla_x f(x;\phi)$ of the discriminator with respect to the input becomes more aggregated over time, as shown in Fig. 3(a). Intuitively, this aggregation may provide fragmented guidance to the generator. We refer to this as the *perceptual narrowing phenomenon*. To quantify this, we devise a metric that counts the connected regions N_{agg} by assigning 1 to gradient values above a threshold and 0 to the rest, followed by a connected component analysis. The ratio of connected regions to the total number of pixels, i.e., $\mathcal{R}_{\text{agg}} = \frac{N_{\text{agg}}}{H \times W}$, indicates the degree of gradient aggregation. Visual representations of connected regions are shown in Fig. 3(a, right).

Extensive experiments on four small-scale datasets using various divergence measures and GAN architectures (detailed in Section 4) consistently showed that the number of connected regions initially increases but then decreases as training progresses, indicating increasing gradient aggregation (Fig. 2b, left). In contrast, with data augmentation or increased data volume to mitigate overfitting, the number of connected regions remains stable (Fig. 2b, middle and right). These findings suggest a link between overfitting and gradient aggregation in limited data GAN training.

Our analysis introduces a fresh perspective through the renormalization group (RG) (Kadanoff, 1966; Wilson, 1971), a technique devised to gradually extract the coarser statistical features of a physical system through local transformation. We observe that, in contrast to the dispersed pattern, the aggregated pattern demonstrates a significant divergence from its coarse-grained counterpart, as illustrated in Fig. 1. This self-dissimilarity (SD) reveals that the system processes information distinctively across different scales (Jacobs & Jacobs, 1992; Wolpert & Macready, 2007), implying that the system is capable of encoding extensive information processing into its dynamics. This suggests that the system is both efficient and possesses a high degree of "plasticity" (learnability) from a neurological perspective (Hensch, 2004). Nonetheless, it also indicates that the system is sensitive and susceptible to minor disturbances that could lead to notable alterations in its dynamics (Achille et al., 2019). These properties hint at the system's tendency toward overfitting and instability, aligning with our empirical findings. Conversely, a minimal SD points to the system's ability to generalize and maintain stability, which is favorable for GAN training.

To verify this point, we employ the Fisher information to examine the properties of deep neural networks during the training process. Although the connection form of weights in neural networks is fixed during training, not all weight connections contribute equally to the final output. We can view $\Psi(x;\phi) \triangleq \nabla_x f(x;\phi)$ as a mapping function parameterized by ϕ, encoding the posterior probability $p(y|x;\phi)$. The input is the generated or real image, and the output is the gradient of the discriminator with respect to the input, $y = \nabla_x f(x;\phi)$. By perturbing the weights and measuring the change in the output distribution using Kullback-Leibler (KL) divergence, we estimate the dependency of the final output on the weights. The second-order Taylor expansion of the KL divergence is:

$$\begin{aligned}&\mathcal{D}_{KL}\left[p(y|x;\phi)||p(y|x;\phi+d\phi)\right]\\ &= \int_{\mathcal{X}} p(y|x;\phi)\log\left[\frac{p(y|x;\phi+d\phi)}{p(y|x;\phi)}\right]dx = \frac{1}{2}d\phi\cdot\mathcal{G}d\phi.\end{aligned} \tag{2}$$

Lemma 3.1. *(Amari, 2016) Any standard f-divergence gives the same Riemannian metric $\mathcal{G}$, which is the Fisher information matrix (FIM)*

$$\mathcal{G} = \mathbb{E}_{x,y}\left[\nabla_\phi \log p(y|x;\phi)\nabla_\phi \log p(y|x;\phi)^\top\right]. \tag{3}$$

The FIM serves as a local measure to assess the effect of weight perturbations on outputs, reflecting the stability of the system $\Psi(x;\phi)$. It also represents the strength of effective connectivity in neural networks and indicates their learnability (Kirkpatrick et al., 2016).

Following Achille et al. (2019), we use the diagonal FIM to reduce the computational cost:

$$tr(\mathcal{G}) = \mathbb{E}_{x,y}\left[\|\nabla_\phi \log p(y|x;\phi)\|^2\right]. \tag{4}$$

As illustrated in Fig. 4(a), Fisher information remains high and increases during the later stages of training, implying decreasing system stability and increasing learnability, hence a higher risk of overfitting and collapse. When the aggregation phenomenon is mitigated through augmentation or increased data volume, Fisher information declines during later training stages, as shown in Fig. 4(b). This aligns with our SD analysis and the CR-GAN approach (Zhang et al., 2020), which uses semantically-preserving augmentation to enhance discriminator robustness by penalizing sensitivity to augmented images.

3.3. Multi-scale Structural Self-dissimilarity

Building upon the previous analysis, we introduce a novel regularization method, termed **m**ulti-**s**cale **s**tructural **s**elf-

dissimilarity (MS^{3}D), designed to enhance the performance of GANs when trained with limited data. This method addresses the perceptual narrowing phenomenon by ensuring that the gradient field $\nabla_x(f(x;\phi))$ of the discriminator $f(x;\phi)$ maintains a consistent pattern across different scales, promoting a more redundant and robust feedback mechanism for the generator.

Formulation. We quantitatively define MS^{3}D using the renormalization group (RG) flow concept. Consider the gradient $\nabla_x(f(x;\phi))$ of the discriminator with respect to the input x, denoted as $\Psi(x;\phi) \triangleq \nabla_x(f(x;\phi))$, where $\Psi(x;\phi)$ is parameterized by ϕ and belongs to the space $\mathcal{F}$ of real-valued functions defined on the Euclidean space $\mathcal{X}$. An RG transformation Γ can be naturally defined for $\Psi(x;\phi)$. In our discrete pixel setting $\mathcal{X}$, Γ can be implemented using transformations like the Kadanoff block-spin transformation (Kadanoff, 1966) or Gaussian filtering. Applying the RG transformation repeatedly results in a coarse-grained function $\Psi(x;\phi)$, generating the evolution of RG, that is, the RG flow.

In the following text, for clarity, let's denote the original system as $\Gamma_{0\to 0}(\Psi(x;\phi)) \triangleq \Psi(x;\phi)$ and the coarse-grained system at scale s as $\Gamma_{0\to s}(\Psi(x;\phi))$. The RG transformation chain is constructed as $\Gamma_{0\to s} = \Gamma_{s-ds\to s} \circ \cdots \circ \Gamma_{0\to ds} \circ \Gamma_{0\to 0}$, where ds represents the scale step, and $\circ$ denotes the composition operator. In other words, $\Gamma_{0\to ds} \circ \Gamma_{0\to 0}(\cdot) = \Gamma_{0\to ds}(\Gamma_{0\to 0}(\cdot))$. For simplicity, we'll use $\Gamma_0 \triangleq \Gamma_{0\to 0}$, $\Gamma_s \triangleq \Gamma_{0\to s}$, and so on. The coarser-grained version of $\Gamma_s(\Psi(x;\phi))$ can be denoted as $\Gamma_{s\to s+ds} \circ \Gamma_s(\Psi(x;\phi))$ or simply as $\Gamma_{s+ds}(\Psi(x;\phi))$. If there is a difference between the coarse-grained system $\Gamma_s(\Psi(x;\phi))$ and its coarser version $\Gamma_{s\to s+ds}(\Gamma_s(\Psi(x;\phi)))$, we can say that the scale s contributes some information processing to the system. To measure this dissimilarity, we introduce self-dissimilarity (SD) as:

$$\begin{aligned}
&\mathcal{D}_{\Gamma_{s\to s+ds}} \\
=&\left|\langle\Gamma_s|\Gamma_{s+ds}\rangle - \frac{1}{2}(\langle\Gamma_s|\Gamma_s\rangle + \langle\Gamma_{s+ds}|\Gamma_{s+ds}\rangle)\right| \\
=&\frac{1}{2}\int_{\mathcal{X}} \left(\Gamma_{s+ds}\left(\Psi(x;\phi)\right) - \Gamma_s(\Psi(x;\phi))\right)^2 dx,
\end{aligned} \tag{5}$$

where $\langle\Gamma_0|\Gamma_1\rangle = \int_{\mathcal{X}} \Gamma_0\Gamma_1 dx$ represents the overlap between two scales. In the context of Kadanoff decimation, SD is calculated by $\frac{1}{2}\left|(\langle\Gamma_s|\Gamma_s\rangle - \langle\Gamma_{s+ds}|\Gamma_{s+ds}\rangle)\right|$ (See the proof in Appendix A). Along the RG flow, the multi-scale structural self-dissimilarity (MS^{3}D) is defined as the integration of SD across scales:

$$\mathcal{D}_{\Gamma_{0\to s}} = \sum_{i=0}^{s/ds} \mathcal{D}_{\Gamma_{i\to i+ds}}. \tag{6}$$

Implementation. We calculate the gradient of the discriminator's logits, denoted as $f(x;\phi)$, with respect to the input image x using the equation $\Psi(x;\phi) = \frac{\partial f(x;\phi)}{\partial x}$. Here, x is an image in the real-valued space $\mathbb{R}^{h\times w\times c}$, where h, w, and c represent the height, width, and number of channels, respectively. The gradient field $\Psi(x;\phi) : \mathbb{R}^{h\times w\times c} \to \mathbb{R}^{h\times w\times c}$ is then transformed into a square matrix $\Psi(x;\phi) : \mathbb{R}^{L\times L} \to \mathbb{R}^{L\times L}$, where $L = \sqrt{h\times w\times c}$. To ensure that the matrix is square, we apply zero-padding to $\Psi(x;\phi)$, resulting in a square matrix $\tilde{\Psi}(x;\phi)$ of size $L \times L$.

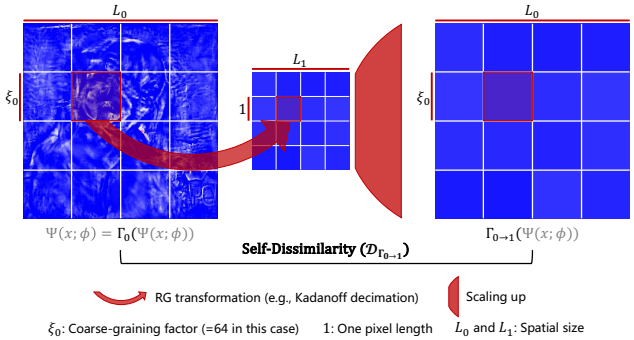


Figure 5. A diagram illustrating the RG transformation process and MS^{3}D computation. It involves iteratively downsampling $\nabla_x f(x;\phi)$ and computing differences before and after downsampling.

For the real function $\tilde{\Psi}(\cdot;\phi)$ and the input image x_0, we normalize the output matrix $\tilde{\Psi}(x_0;\phi)$ to the range $[0,1]$:

$$\tilde{\Psi}(x;\phi) \leftarrow \left|\frac{\tilde{\Psi}(x_0;\phi)}{\max(|\tilde{\Psi}(x_0;\phi)|)}\right|. \tag{7}$$

We then use the Kadanoff block-spin transformation to coarse-grain the system, as illustrated in Fig. 5. Although the Kadanoff decimation is the simplest renormalization group (RG) transformation, it yields meaningful and robust outcomes, as verified in our ablation study (detailed in Section 4). During the transformation $\Gamma_{s\to s+ds}$, the matrix $\Gamma_s(\tilde{\Psi}(x_0;\phi))$, with spatial dimensions $L_s \times L_s$, is tiled by blocks of size $\zeta_s \times \zeta_s$, where ζ is the coarse-graining factor. In this paper, we set $\zeta_0 = \zeta_1 = \cdots = \zeta_s = 2$. The coarse-grained matrix $\Gamma_{s+ds}(\tilde{\Psi}(x_0;\phi))$ is obtained by replacing each block with its average value. This matrix retains the original spatial dimensions $L_{s+ds} = L_s$, but the number of elements is reduced by a factor of $(\zeta_s)^2$. Mathematically, this process is represented as follows:

$$\begin{aligned}
&\left\{\Gamma_{s+ds}(\tilde{\Psi}(x_0;\phi))\right\}(i,j) \\
=&\frac{1}{(\zeta_s)^2}\sum_{m=0}^{\zeta_s-1}\sum_{n=0}^{\zeta_s-1}\left\{\Gamma_s(\tilde{\Psi}(x_0;\phi))\right\}\left(\lfloor\frac{i}{\zeta_s}\rfloor\cdot\zeta_s + m,\right. \\
&\left.\lfloor\frac{j}{\zeta_s}\rfloor\cdot\zeta_s + n\right),
\end{aligned} \tag{8}$$

where $\lfloor\cdot\rfloor$ is the floor function, and $\{\Gamma(\cdot)\}(i,j)$ indicates the element at the i-th row and j-th column of $\Gamma(\cdot)$.

Table 1. Performance comparison of various GAN loss functions (divergence measures). ⇑ denotes that higher values are preferable, while ⇓ indicates that lower values are better.

METHOD	OXFORDDOG		FFHQ-2.5K	
	FID ⇓	IS ⇑	FID ⇓	IS ⇑
STYLEGAN2 (NS)	64.26	9.69	48.11	3.50
+ WASSERSTEIN	82.18	9.94	38.64	4.11
+ LS	216.42	2.69	213.93	2.17
+ RAHINGE	38.68	11.14	32.34	3.96
+ MS^{3}D	47.05	10.73	33.46	4.61
+ MS^{3}D + RAHINGE	**37.71**	**12.54**	**31.95**	**4.70**

Subsequently, the matrix $\Gamma_{s+ds}(\tilde{\Psi}(x_0;\phi))$ is tiled by blocks of size $\zeta^{s+ds} \times \zeta^{s+ds}$, and its coarser-grained version $\Gamma_{s+2ds}(\tilde{\Psi}(x_0;\phi))$ is obtained by replacing each block with its average value. This process is repeated until $(\zeta)^{t+1} > L$, i.e., the coarse-grained transformation cannot be applied anymore, where t is the number of iterations and $1 \leq t \leq \log_\zeta L$. For our purposes, we set $\zeta = 2$. To integrate the MS^{3}D regularization within widely-used computational frameworks like PyTorch (Paszke et al., 2019) and TensorFlow (Abadi et al., 2016), we utilize a convolution operation with an average filter of size 2×2 and a stride of 2 to perform the Kadanoff decimation. The coarse-grained output is then scaled back to the original dimensions by a factor of 2. In Appendix B, we present the PyTorch-like pseudo-code of MS^{3}D calculation. The dissimilarity between the two versions is computed by the overlap of the current version and the coarser-grained version (Eq. (5)). We simply implement SD as the mean squared error (MSE). Hence, the MS^{3}D is the cumulative sum of SDs across all scales, formulated as:

$$\mathcal{D}_{\Gamma_{0\to t}} = \sum_{i=0}^{t} \|\Gamma_i(\Psi(x_0;\phi)) - \Gamma_{i\to i+1}(\Psi(x_0;\phi))\|_2^2. \quad (9)$$

The loss function for the discriminator is:

$$\begin{aligned}\mathcal{L}_{\text{dis}} = &- \mathbb{E}_x \log f(x;\phi) - \mathbb{E}_z \log(1 - f(g(z));\phi) \\ &+ \lambda \mathbb{E}_x \mathcal{D}_{\Gamma_{0\to t}}(\nabla_x(f(x;\phi))),\end{aligned} \quad (10)$$

where λ is the weight of the MS^{3}D regularization term.

4. Experiments

4.1. Experimental Settings

Datasets. Our experiments utilize four diverse datasets: OxfordDog (from Oxford-IIIT pet dataset (Parkhi et al., 2012), detailed in Appendix C), Flickr-Faces-HQ (FFHQ) (Karras et al., 2019), MetFaces (Karras et al., 2020a), and BreCaHAD (Aksac et al., 2019). Detailed descriptions of these datasets are provided in Appendix C.

Metrics. We employ Inception Score (IS) (Salimans et al., 2016), Fréchet Inception Distance (FID) (Heusel et al., 2017), and Kernel Inception Distance (KID) (Binkowski et al., 2018) to evaluate our models. We utilize the official implementations of these metrics as provided by Karras et al. (2020b).

Table 2. Comparison of techniques for mitigating discriminator overfitting under limited data.

METHOD	OXFORDDOG		FFHQ-2.5K	
	FID ⇓	IS ⇑	FID ⇓	IS ⇑
STYLEGAN2	64.26	9.69	48.11	3.50
+ LECAM	102.87	8.02	68.85	3.53
+ DIGGAN	61.84	10.07	48.45	3.68
+ KD-DLGAN	54.06	9.73	49.31	3.67
+ MS^{3}D	47.05	10.73	33.46	4.61
+ CR + ADA	29.81	11.76	**19.98**	4.53
+ MS^{3}D + ADA	**25.94**	**12.08**	20.23	**4.71**

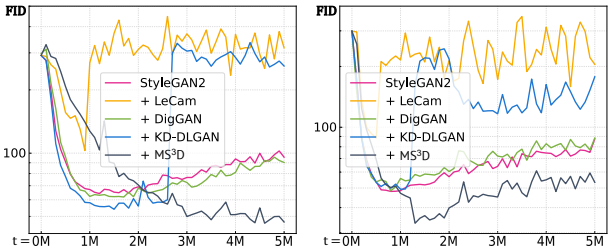
Table 3. Comparison of various GAN architectures and techniques.

METHOD		FID ⇓	IS ⇑
DCGAN	VANILLA	107.33	3.23
	+ AUXILIARY ROTATIONS	103.43	2.85
	+ DROPOUT	77.35	4.11
	+ NOISE	86.13	**5.40**
	+ MS^{3}D	**59.37**	4.39
SNGAN	VANILLA	57.08	3.42
	+ AUXILIARY ROTATIONS	54.85	3.63
	+ DROPOUT	48.54	4.02
	+ NOISE	55.99	3.33
	+ MS^{3}D	**47.87**	**4.18**

4.2. Results: Various Limited Data Settings

In Table 1, we compare our method with multiple GAN losses, including Wasserstein (Arjovsky et al., 2017), Least Squares (LS) (Mao et al., 2017), RaHinge (Jolicoeur-Martineau, 2019), and NS (non-saturated, vanilla StyleGAN2) (Goodfellow et al., 2014). We observe that the state-of-the-art divergence metrics, such as Wasserstein and LS, which are originally devised to stabilize GAN training, perform poorly in limited data settings. Surprisingly, RaHinge loss performs very well, unlike the poor results reported by Tseng et al. (2021), and it outperforms our default setting, i.e., StyleGAN2 + MS^{3}D (abbreviated as + MS^{3}D). A possible reason is that the relativistic discriminator reduces the probability of real data being classified as real, thus enhancing the discriminator's generalization ability. Nevertheless, our method still surpasses it and achieves the best performance on + MS^{3}D + RaHinge. This indicates that our method can effectively integrate with other methods.

In Table 2, we show the comparison with other regularization methods, namely: consistency regularization (CR) (Zhang et al., 2020), exponential moving average regularization (LeCam) (Tseng et al., 2021), discriminator gradient gap regularization (DigGAN) (Fang et al., 2022), and knowledge distillation (KD-DLGAN) (Cui et al., 2023). CR aims to constrain the discriminator's sensitivity to augmented data, relying on augmentation, while our method is



(a) OxfordDog (b) FFHQ-2.5K

Figure 6. Evolution of FID values on OxfordDog and FFHQ-2.5K datasets over various training steps.

Table 4. Quantitative evaluation (FID ⇓) on varying data scales of the FFHQ dataset.

METHOD	1K	2K	4K	6K	8K	10K	70K
STYLEGAN2	87.86	54.96	27.63	22.97	15.98	12.50	6.50
+ LECAM	99.90	88.51	69.55	33.46	16.06	**10.48**	6.53
+ DIGGAN	81.42	56.33	31.26	20.92	15.77	12.61	6.63
+ KD-DLGAN	83.67	55.34	29.91	**19.84**	15.99	12.73	7.11
+ RAHINGE	93.14	56.64	25.72	20.43	**12.55**	10.98	**6.10**
+ MS^{3}D	**81.30**	**49.17**	**24.17**	19.86	12.61	10.59	6.28

augmentation-free. To provide a comprehensive comparison, we also combine adaptive discriminator augmentation (ADA) (Karras et al., 2020a) with our method. As shown in Table 2, our method outperforms LeCam, DigGAN, and KD-DLGAN, which are specifically designed for limited data scenarios. Additionally, in combination with ADA (+ MS^{3}D + ADA), our method surpasses CR with ADA (+ CR + ADA).

Furthermore, we compare our method across different GAN architectures, including DCGAN (Radford et al., 2016) and SNGAN (Miyato et al., 2018), and various improvement techniques such as auxiliary rotation prediction (Chen et al., 2019), dropout (Srivastava et al., 2014), and noise injection (Sønderby et al., 2017). As demonstrated in Table 3, our method performs well not only with StyleGAN architectures but also with traditional convolution-based DCGAN and ResNet-based SNGAN, indicating that our method's effectiveness is not dependent on the underlying neural network structure, likely due to its primary impact on gradients.

To further demonstrate the effectiveness of our method in improving the quality of generated images under limited data conditions, we present FID values at various data scales in Table 4. Our method significantly reduces FID values, particularly at more limited data scales such as 1K, 2K and 4K. In Fig. 6, we illustrate the FID values over training steps for OxfordDog and FFHQ-2.5K. These results confirm that our method effectively mitigates the image quality deterioration typically observed with limited data. The generated images, as depicted in Fig. 10, visibly demonstrate that our method yields more realistic images in comparison to baseline models.

Table 5. Quantitative comparison on small datasets, MetFaces and BreCaHAD.

METHOD	METFACES			BRECAHAD		
	FID ⇓	IS ⇑	KID ⇓	FID ⇓	IS ⇑	KID ⇓
STYLEGAN2	53.21	3.16	0.028	97.06	**3.10**	0.095
+ LECAM	56.67	2.43	0.102	83.74	2.51	0.046
+ DIGGAN	53.97	3.09	0.031	105.45	3.05	0.095
+ KD-DLGAN	54.05	3.11	0.029	86.25	3.26	0.083
+ MS^{3}D	37.71	3.85	0.017	67.12	2.62	0.038
+ CR + ADA	29.91	4.20	0.008	22.69	2.81	0.011
+ MS^{3}D + ADA	23.09	4.15	**0.004**	**20.22**	2.81	**0.006**
+ MS^{3}D + ADA + CR	**21.54**	**4.28**	**0.004**	21.09	2.92	0.007

Table 6. Quantitative results of transfer learning with Freeze-D technique applied to MetFaces dataset. The model is pretrained on FFHQ-70K.

METHOD	METFACES		
	FID ⇓	IS ⇑	KID ⇓
STYLEGAN2	53.21	3.16	0.028
+ FREEZE-D	47.74	3.32	0.020
+ MS^{3}D + FREEZE-D	**36.95**	**3.56**	**0.013**

4.3. Results: Small Dataset

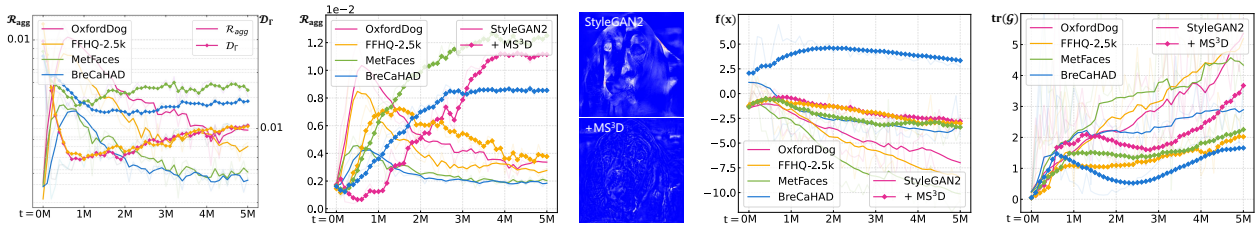
We also conducted experiments on small datasets, MetFaces and BreCaHAD, which contain only 1,203 and 1,750 training images, respectively. Table 5 compares our method with other techniques on these datasets, demonstrating that our method achieves the best performance. Following Karras et al. (2020a), we use KID for quantitative evaluation due to its stability on small datasets. The generated images in Fig. 10 show that our method produces more realistic images than baseline models.

4.4. Results: Transfer Learning

To cope with the difficulty of data collection, recent researchers have leveraged transfer learning in the image generation setting. Wang et al. (2020) use fine-tuning to transfer the knowledge of models pre-trained on external large-scale datasets. Some works propose to fine-tune only part of the model, such as Mo et al. (2020) propose to fix the highest resolution layer of the discriminator (Freeze-D) for knowledge transfer, showing competitive results. In Table 6, we implement this technique and show the results. We find that our method achieves consistent gains on all datasets, although the gains are very marginal.

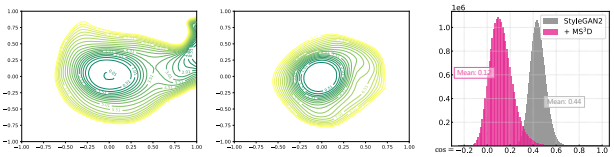
4.5. Analysis and Ablation Studies

Training dynamics. In Section 3.2, we demonstrate that perceptual narrowing forms a feedback system $\Psi(x;\phi)$ characterized by high learnability and sensitivity, exhibiting significant self-dissimilarity (SD). Thus, we define multi-scale structural self-dissimilarity (MS^{3}D) and incorporate it to enhance the performance of GANs with limited data. To validate the efficacy of MS^{3}D regularization ($\mathcal{D}_\Gamma$), we illustrate



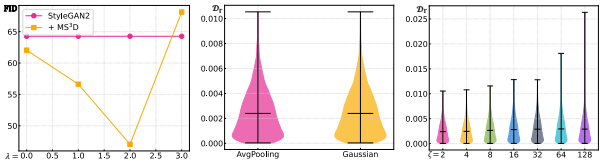
(a) (b) (c) (d) (e)

Figure 7. (a) The relationship between perceptual narrowing and MS^{3}D. (b) Reduction in gradient aggregation post-MS^{3}D application. (c) Visualization of gradients post-MS^{3}D, showing a dispersed pattern similar to augmentation methods. (d) Mitigation of discriminator overfitting to validation data with MS^{3}D. (e) Decrease in Fisher information post-MS^{3}D, indicating enhanced stability in training dynamics.



(a) Loss landscapes (b) Similarities

Figure 8. (a) Loss landscapes of the discriminator without MS^{3}D (left) and with MS^{3}D (right). (b) Average cosine similarity of the discriminator's data embeddings.



(a) Parameter λ (b) Filter Γ (c) Factor ζ

Figure 9. Results of the ablation study. (a) Ablation study on the hyper-parameter λ. (b) Ablation of the RG transformation filter. (c) Ablation of the coarse-graining factor.

Table 7. Computational complexity of different methods.

METHOD	TIME (S / KIMG)	CPU MEMORY (G)	GPU MEMORY (G)
STYLEGAN2	**9.73**	**5.354**	9.127
+ MS^{3}D	9.75	5.374	9.127

Table 8. Ablation study on MS^{3}D regularization applied to real and generated data.

CONFIGURATION $\mathcal{D}_\Gamma(\Psi(x))$		OXFORDDOG	FFHQ-2.5K
x = REAL	x = GENERATED		
✗	✓	62.40	86.51
✓	✗	47.05	33.46
✓	✓	**42.73**	**36.31**

the relationship between perceptual narrowing and MS^{3}D in Fig. 7(a), revealing their close connection. Fig. 7(b) displays the gradient patterns post MS^{3}D application, showing a more dispersed pattern similar to augmentation methods, which is visualized in Fig. 7(c). In Fig. 7(d), validation data demonstrate our method's superior generalization, evidenced by the discriminator's lower confidence in evaluating validation data. Moreover, with minimal training data, our method ensures stable training without collapsing. Fig. 7(e) shows that after applying MS^{3}D, the Fisher information of $\Psi(x;\phi)$ decreases, indicating more stable training dynamics.

Loss landscape. In addition to examining the impact of MS^{3}D on GAN learning dynamics from the RG perspective, we also assess it from the loss landscape perspective. Figs. 8(a) and (b) reveal how MS^{3}D influences the discriminator's loss landscape under limited data conditions, making the loss landscape flatter. A flatter loss landscape generally implies enhanced model generalization and training stability (Li et al., 2018).

Embedding space. Further observations show that the mean cosine similarity of the discriminator's data embeddings significantly increases by the end of training, compared to the initial phase. This increase indicates a narrower embedding space for the discriminator, suggesting a large gap between real and generated image representations. This discrepancy can enable the discriminator to easily achieve perfect true-false classification, leading to overfitting. Conversely, the application of MS^{3}D or data augmentation prevents this narrowing, maintaining low average cosine similarity levels, as depicted in Fig. 8(c). These findings imply that perceptual narrowing may cause a narrower embedding space, and MS^{3}D can effectively mitigate this issue, thereby enhancing discriminator generalization. Additional details are available in Appendix A.

Computational complexity. Our RG process involves iteratively downsampling $\Psi(x;\phi)$ and computing differences pre- and post-downsampling. Using pooling operators in PyTorch, this procedure does not require trainable parameters. Table 7 summarizes the computational overhead and memory usage during training on OxfordDog with Nvidia RTX 4090, indicating negligible resource demands.

Ablation study. Fig. 9 (a) presents results for different λ

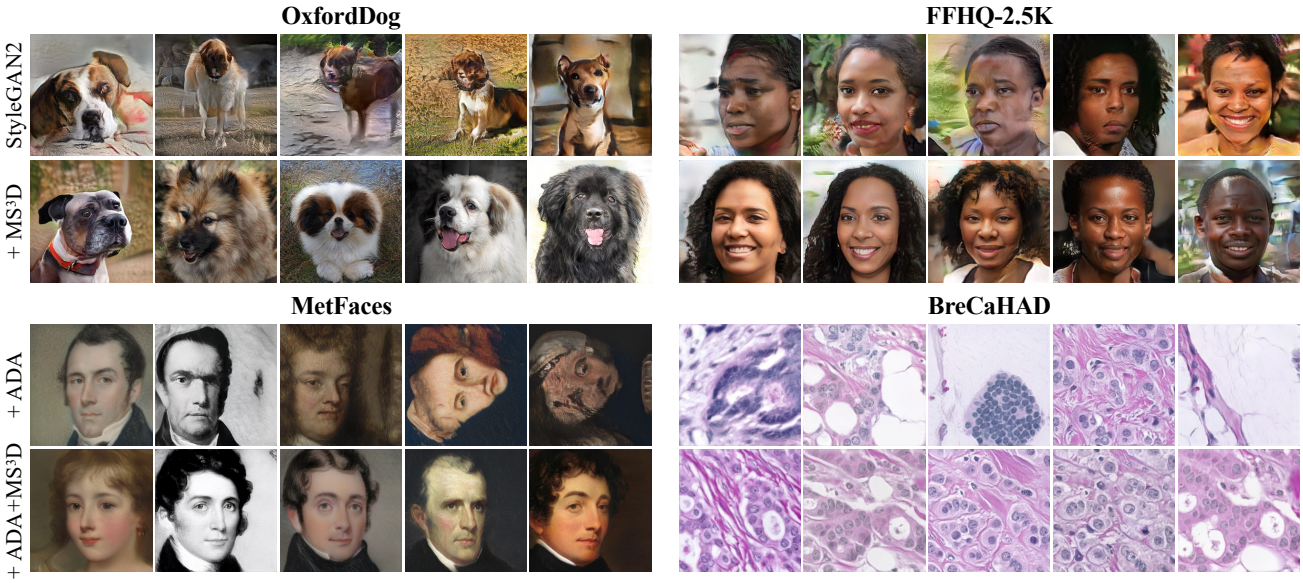


Figure 10. Example generated images for several datasets with limited amount of training data.

Table 9. Quantitative results on SNGAN with FFHQ-2.5K.

METHOD	FID ↓	IS ↑
SNGAN	**45.31**	4.32
+ MS^{3}D	46.14	**4.37**

values in Eq. (10), showing optimal performance at $\lambda = 10$. Larger λ values may interfere with other constraint terms in StyleGAN2, which has multiple constraints.

We also conduct ablation experiments on different RG transformations, introducing a Gaussian kernel as an alternative to the Kadanoff block-spin transformation (Fig. 9(b)). The results show minimal differences, leading us to adopt the simpler method. In Eq. (8), we chose a coarse-graining factor of 2, a common practice, and varied this factor in numerical experiments on OxfordDog (Fig. 9(c)), noting that larger factors increased variance and reduced numerical stability.

For $\mathcal{D}_{\Gamma_{0\to t}}(\Psi(x;\phi))$ in Eq. (10), x can be either real or generated data. Table 8 compares results on OxfordDog and FFHQ-2.5K for real and generated data, finding optimal performance when x includes both. However, regularization on real data alone is more effective in avoiding discriminator overfitting, leading us to choose this for computational efficiency.

Limitations. For low-resolution image generation models (e.g., 32×32), such as DCGAN and SNGAN, our method may not yield significant improvements. These models often exhibit stable training dynamics due to their reduced internal differences. Experiments conducted on SNGAN with FFHQ-2.5K, which intentionally minimizes internal differences (details in Appendix C), support this observation. The results, presented in Table 9, confirm that GANs trained on such datasets tend to be relatively easy to stabilize. Furthermore, the influence of our method on gradients, which promotes more uniform and dispersed gradients, may not be particularly beneficial for low-resolution image generation.

5. Conclusion and Future Work

In our study, we are analyzing the gradient pattern that the generator receives from the discriminator at various scales based RG flow. In particular, we are interested in the relationship between the gradient pattern and discriminator deterioration in the limited data setting. Based on this, we propose a new regularization method, called MS^{3}D. We conduct extensive experiments on various datasets and tasks, and the results show that our method can significantly improve the performance of GANs under limited data. Recently, RG theory has been linked with the notions of information theory (Gordon et al., 2021). Therefore, an intriguing future direction is to explore the connection between the gradient pattern and the perceptual narrowing phenomenon from the angle of information theory.

Acknowledgements

This work is supported by the Fundamental Research Funds for the Central Universities under Grant 1082204112364.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I. J., Harp, A., Irving, G., Isard, M., Jia, Y., Józefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D. G., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P. A., Vanhoucke, V., Vasudevan, V., Viégas, F. B., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *CoRR*, abs/1603.04467, 2016.

Achille, A., Rovere, M., and Soatto, S. Critical learning periods in deep networks. In *7th International Conference on Learning Representations*, 2019.

Aksac, A., Demetrick, D. J., Ozyer, T., and Alhajj, R. Brecahad: a dataset for breast cancer histopathological annotation and diagnosis. *BMC Research Notes*, Dec 2019.

Amari, S.-i. *Information geometry and its applications*, volume 194. Springer, 2016.

Arjovsky, M. and Bottou, L. Towards principled methods for training generative adversarial networks. In *5th International Conference on Learning Representations*, 2017.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pp. 214–223, 2017.

Bagrov, A. A., Iakovlev, I. A., Iliasov, A. A., Katsnelson, M. I., and Mazurenko, V. V. Multiscale structural complexity of natural patterns. *Proceedings of the National Academy of Sciences*, 117(48):30241–30251, 2020.

Binkowski, M., Sutherland, D. J., Arbel, M., and Gretton, A. Demystifying MMD gans. In *6th International Conference on Learning Representations*, 2018.

Bradde, S. and Bialek, W. Pca meets rg. *Journal of statistical physics*, 167:462–475, 2017.

Brock, A., Donahue, J., and Simonyan, K. Large scale GAN training for high fidelity natural image synthesis. In *7th International Conference on Learning Representations*, 2019.

Cam, L. L. *Asymptotic Methods in Statistical Decision Theory*. Springer, Jan 1986.

Chen, T., Zhai, X., Ritter, M., Lucic, M., and Houlsby, N. Self-supervised gans via auxiliary rotation loss. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 12154–12163, 2019.

Cui, K., Huang, J., Luo, Z., Zhang, G., Zhan, F., and Lu, S. Genco: Generative co-training for generative adversarial networks with limited data. In *Thirty-Sixth AAAI Conference on Artificial Intelligence*, pp. 499–507, 2022.

Cui, K., Yu, Y., Zhan, F., Liao, S., Lu, S., and Xing, E. P. KD-DLGAN: data limited image generation via knowledge distillation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3872–3882, 2023.

Denton, E. L., Chintala, S., Szlam, A., and Fergus, R. Deep generative image models using a laplacian pyramid of adversarial networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015*, pp. 1486–1494, 2015.

Fang, T., Sun, R., and Schwing, A. G. Diggan: Discriminator gradient gap regularization for GAN training with limited data. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022*, volume 35, pp. 31782–31795, 2022.

Fisher, M. E. Renormalization group theory: Its basis and formulation in statistical physics. *Reviews of Modern Physics*, 70(2):653, 1998.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. C., and Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014*, pp. 2672–2680, 2014.

Gordon, A., Banerjee, A., Koch-Janusz, M., and Ringel, Z. Relevance in the renormalization group and in information theory. *Physical Review Letters*, 126(24):240601, 2021.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, pp. 5767–5777, 2017.

Gulrajani, I., Raffel, C., and Metz, L. Towards GAN benchmarks which require generalization. In *7th International Conference on Learning Representations*, 2019.

Hensch, T. K. Critical period regulation. *Annual Review of Neuroscience*, pp. 549–579, Jul 2004.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, pp. 6626–6637, 2017.

Jacobs, K. and Jacobs, K. Elements of information theory. *Discrete Stochastics*, pp. 155–183, 1992.

Jenni, S. and Favaro, P. On stabilizing generative adversarial training with noise. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 12145–12153, 2019.

Jiang, L., Dai, B., Wu, W., and Loy, C. C. Deceive D: adaptive pseudo augmentation for GAN training with limited data. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021*, pp. 21655–21667, 2021.

Jolicoeur-Martineau, A. The relativistic discriminator: a key element missing from standard GAN. In *7th International Conference on Learning Representations*, 2019.

Kadanoff, L. P. Scaling laws for ising models near t c. *Physics Physique Fizika*, 2(6):263, 1966.

Karras, T., Aila, T., Laine, S., and Lehtinen, J. Progressive growing of gans for improved quality, stability, and variation. In *6th International Conference on Learning Representations*, 2018.

Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410, 2019.

Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., and Aila, T. Training generative adversarial networks with limited data. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020*, 2020a.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. Analyzing and improving the image quality of stylegan. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8107–8116, 2020b.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N. C., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., and Hadsell, R. Overcoming catastrophic forgetting in neural networks. *CoRR*, abs/1612.00796, 2016.

Koch-Janusz, M. and Ringel, Z. Mutual information, neural networks and the renormalization group. *Nature Physics*, 14(6):578–582, 2018.

Kumari, N., Zhang, R., Shechtman, E., and Zhu, J. Ensembling off-the-shelf models for GAN training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10641–10652, 2022.

Li, H., Xu, Z., Taylor, G., Studer, C., and Goldstein, T. Visualizing the loss landscape of neural nets. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018*, pp. 6391–6401, 2018.

Lin, H. W., Tegmark, M., and Rolnick, D. Why does deep and cheap learning work so well. *Journal of Statistical Physics*, 168(6):1223–1247, Sep 2017.

Liu, S., Wang, T., Bau, D., Zhu, J., and Torralba, A. Diverse image generation via self-conditioned gans. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14274–14283, 2020.

Machacek, M. E. and Vaughn, M. T. Two-loop renormalization group equations in a general quantum field theory:(iii). scalar quartic couplings. *Nuclear Physics B*, 249 (1):70–92, 1985.

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

Mehta, P. and Schwab, D. J. An exact mapping between the variational renormalization group and deep learning. *CoRR*, abs/1410.3831, 2014.

Mescheder, L. M., Geiger, A., and Nowozin, S. Which training methods for gans do actually converge? In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 3478–3487, 2018.

Miyato, T. and Koyama, M. cgans with projection discriminator. In *6th International Conference on Learning Representations*, 2018.

Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. Spectral normalization for generative adversarial networks. In *6th International Conference on Learning Representations*, 2018.

Mo, S., Cho, M., and Shin, J. Freeze discriminator: A simple baseline for fine-tuning gans. *CoRR*, abs/2002.10964, 2020.

Parkhi, O. M., Vedaldi, A., Zisserman, A., and Jawahar, C. V. Cats and dogs. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3498–3505, 2012.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E. Z., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library. In *Advances in*

Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, pp. 8024–8035, 2019.

Radford, A., Metz, L., and Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. In *4th International Conference on Learning Representations*, 2016.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763, 2021.

Salimans, T., Goodfellow, I. J., Zaremba, W., Cheung, V., Radford, A., and Chen, X. Improved techniques for training gans. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016*, pp. 2226–2234, 2016.

Shorten, C. and Khoshgoftaar, T. M. A survey on image data augmentation for deep learning. *J. Big Data*, 6:60, 2019.

Sønderby, C. K., Caballero, J., Theis, L., Shi, W., and Huszár, F. Amortised MAP inference for image super-resolution. In *5th International Conference on Learning Representations*, 2017.

Song, J. and Ermon, S. Bridging the gap between f-gans and wasserstein gans. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 9078–9087, 2020.

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958, 2014.

Stanley, H. E. Scaling, universality, and renormalization: Three pillars of modern critical phenomena. *Reviews of modern physics*, 71(2):S358, 1999.

Tran, N., Tran, V., Nguyen, N., Nguyen, T., and Cheung, N. Towards good practices for data augmentation in GAN training. *CoRR*, abs/2006.05338, 2020.

Tran, N., Tran, V., Nguyen, N., Nguyen, T., and Cheung, N. On data augmentation for GAN training. *IEEE Trans. Image Process.*, 30:1882–1897, 2021.

Tseng, H., Jiang, L., Liu, C., Yang, M., and Yang, W. Regularizing generative adversarial networks under limited data. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7921–7931, 2021.

Wang, Y., Gonzalez-Garcia, A., Berga, D., Herranz, L., Khan, F. S., and van de Weijer, J. Minegan: Effective knowledge transfer from gans to target domains with few images. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9329–9338, 2020.

Webster, R., Rabin, J., Simon, L., and Jurie, F. Detecting overfitting of deep generative networks via latent recovery. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 11273–11282, 2019.

Wilson, K. G. Renormalization group and critical phenomena. i. renormalization group and the kadanoff scaling picture. *Physical review B*, 4(9):3174, 1971.

Wolpert, D. H. and Macready, W. G. Using self-dissimilarity to quantify complexity. *Complex.*, 12(3):77–85, 2007.

Wolpert, D. H. and Macready, W. G. Self-dissimilarity: An empirically observable complexity measure. In *Unifying Themes In Complex Systems, Volume 1*, pp. 625–644. CRC Press, 2018.

Zhang, H., Xu, T., and Li, H. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *IEEE International Conference on Computer Vision*, pp. 5908–5916, 2017.

Zhang, H., Goodfellow, I. J., Metaxas, D. N., and Odena, A. Self-attention generative adversarial networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 7354–7363, 2019.

Zhang, H., Zhang, Z., Odena, A., and Lee, H. Consistency regularization for generative adversarial networks. In *8th International Conference on Learning Representations*. OpenReview.net, 2020.

Zhao, J. J., Mathieu, M., and LeCun, Y. Energy-based generative adversarial networks. In *5th International Conference on Learning Representations*, 2017.

Zhao, S., Liu, Z., Lin, J., Zhu, J., and Han, S. Differentiable augmentation for data-efficient GAN training. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020*, 2020a.

Zhao, Z., Zhang, Z., Chen, T., Singh, S., and Zhang, H. Image augmentations for GAN training. *CoRR*, abs/2006.02595, 2020b.

Zhao, Z., Singh, S., Lee, H., Zhang, Z., Odena, A., and Zhang, H. Improved consistency regularization for gans. In *Thirty-Fifth AAAI Conference on Artificial Intelligence*, pp. 11033–11041, 2021.

A. Discussion

A.1. Simple Proof of SD Computation

In this section, we demonstrate the proof for computing the structural self-dissimilarity (SD) as presented in the main text. We begin by defining SD as follows:

$$\mathcal{D}_{\Gamma_{s\to s+ds}} = \left| \langle \Gamma_s | \Gamma_{s+ds} \rangle - \frac{1}{2}(\langle \Gamma_s | \Gamma_s \rangle + \langle \Gamma_{s+ds} | \Gamma_{s+ds} \rangle) \right| . \tag{11}$$

In the main text, we showed that SD can be computed using the equation:

$$\mathcal{D}_{\Gamma_{s\to s+ds}} = \frac{1}{2} \left| (\langle \Gamma_s | \Gamma_s \rangle - \langle \Gamma_{s+ds} | \Gamma_{s+ds} \rangle) \right| . \tag{12}$$

proof: Consider the Kadanoff block-spin transformation:

$$\{\Gamma_{s+ds}(\Psi(x;\phi))\}(i,j) = \frac{1}{(\zeta_s)^2} \sum_{m=0}^{\zeta_s-1} \sum_{n=0}^{\zeta_s-1} \{\Gamma_s(\Psi(x;\phi))\} \left(\lfloor \frac{i}{\zeta_s} \rfloor \cdot \zeta_s + m, \lfloor \frac{j}{\zeta_s} \rfloor \cdot \zeta_s + n \right) . \tag{13}$$

Thus, we have:

$$\begin{aligned}
\langle \Gamma_s | \Gamma_{s+ds} \rangle &= \frac{1}{(L_{s+ds})^2} \sum_{i=0}^{L_s-1} \sum_{j=0}^{L_s-1} \Gamma_s \{(\Psi(x;\phi))\}(i,j) \sum_{m=0}^{\zeta_s-1} \sum_{n=0}^{\zeta_s-1} \{\Gamma_s(\Psi(x;\phi))\} \left(\lfloor \frac{i}{\zeta_s} \rfloor \cdot \zeta_s + m, \lfloor \frac{j}{\zeta_s} \rfloor \cdot \zeta_s + n \right) \\
&= \frac{(\zeta_s)^2}{(L_{s+ds})^2} \sum_{i=0}^{L_s-1} \sum_{j=0}^{L_s-1} \left(\Gamma_s \{(\Psi(x;\phi))\}(i,j) \right)^2 \\
&= \frac{(\zeta_s)^2}{(L_{s+ds})^2} \cdot (L_s)^2 (\Gamma_s(\Psi(x;\phi)))^2 \\
&= \langle \Gamma_s | \Gamma_s \rangle
\end{aligned} \tag{14}$$

A.2. Comparison with ADA

While adaptive discriminator augmentation (ADA) (Karras et al., 2020a) is an effective method for training GANs under limited data, we observe that with very small datasets, ADA increases the probability of augmentation and leads to "augmentations leak". For instance, faces generated on MetFaces may show rotation (see Fig. 10 and 18). In Fig. 11, we compare ADA in terms of the degree of aggregation of the gradients. Our approach demonstrates better performance in avoiding aggregated patterns on small datasets like MetFaces (Karras et al., 2020a) and BreCaHAD (Aksac et al., 2019).

A.3. Cosine Similarity of the Data Embedding Space

We use the discriminator as a feature extractor and obtain data representations through forward inference. We then compute the pairwise cosine similarity of the representations for all real images. Observing the training process, the average cosine similarity of the representations increases significantly, as shown in Fig. 12. A higher average cosine similarity indicates a narrowing of the representation space. Considering the surface area ratio of the unit hypersphere: in two dimensions, $\arccos(0.56) = 63.90°$, implying that a cosine similarity of 0.56 can "occupy"

$$\frac{63.90°}{360°} = 17.75\%$$

of the 2D unit circle. In three dimensions, it is only 6.96%. This low proportion suggests a sparse distribution of points with a cosine similarity of 0.56 in high-dimensional space, possibly explaining the narrowing of the representation space.

A.4. Landscape of Discriminator

In Fig. 14, we present the loss landscape of the discriminator without MS^{3}D, with MS^{3}D, and with ADA. We find that our MS^{3}D contributes to a flatter loss landscape. A flatter loss landscape often correlates with higher model generalization and training stability (Li et al., 2018).

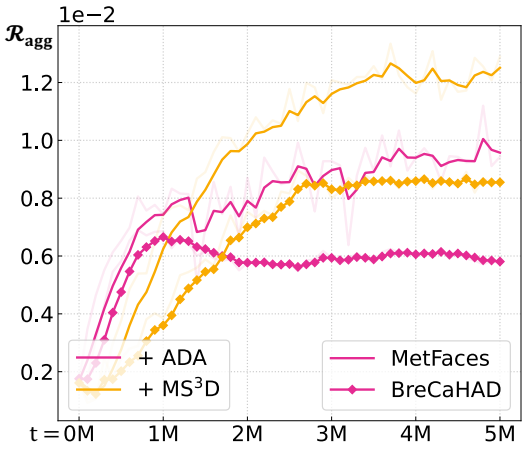


Figure 11. Degree of aggregation of the gradients.

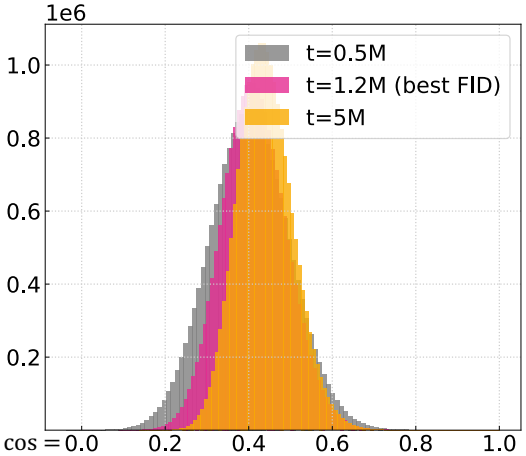


Figure 12. Cosine similarity over steps.

A.5. Pattern of $\nabla_x f(x;\phi)$

Additional visualizations of $\nabla_x f(x;\phi)$ are shown in Fig. 15 and 16. In Fig. 17, we compute the multi-scale structural self-dissimilarity (MS^{3}D) of the discriminator's activation maps, revealing that MS^{3}D effectively captures the complexity of natural patterns.

B. Illustration of MS^{3}D

Pseudocode. In this section, we provide the pseudocode of the proposed multi-scale structural self-dissimilarity (MS^{3}D) regularization. The pseudocode is shown in Fig 13.

```
x  # real images
G  # generator
D  # discriminator

# Minimize logits for generated images
logits = D(G(x))
loss_d_adv_fake = F.softplus(logits)

# Maximize logits for real images
logits = D(x)
loss_d_adv_real = F.softplus(-logits)

# MS3D regularization
grads = torch.autograd.grad(outputs=logits, inputs=x)
grads = normalize(grads)
grads = reshape_to_square(grads)
spatial_size = get_spatial_size(grads)
ms3d = 0
for i in range(torch.log2(spatial_size)):
    grads_down = F.avg_pool2d(grads, kernel_size=2, stride=2)  # Kadanoff decimation
    # or use: grads_down = F.conv2d(grads, guass_kernel, stride=2)
    grads_recn = F.interpolate(grads_down, size=grads.shape[2:], mode='nearest')
    ms3d += F.mse_loss(grads, grads_recn)
    grads = grads_down
loss_d = loss_d_adv_fake + loss_d_adv_real + ms3d
loss_d.backward()
```

Figure 13. Pseudocode of the proposed multi-scale structural self-dissimilarity (MS^{3}D) regularization.

C. Experimental settings

C.1. Datasets

OxfordDog. The Oxford-IIIT pet dataset (Parkhi et al., 2012) contains images of cat and dog species. For this study, we select only dog images, apply a dog face detection model to crop and resize them to a uniform square format of 256×256 pixels, centering the dog's face as much as possible. We randomly select 4,492 images for training and use the remaining 498 images as the test set. This dataset includes a wide variety of dog breeds (about 25 dog breeds), poses, and backgrounds, thus providing a challenging testbed for our experiments. We refer to this dataset as OxfordDog.

FFHQ. The Flickr-Faces-HQ (FFHQ) dataset (Karras et al., 2019) includes 70,000 high-quality images of human faces, covering a wide range of ages, ethnicities, and image backgrounds, as well as various accessories like eyeglasses and hats. Each image was sourced from Flickr, aligned, and cropped to ensure consistency and quality.

MetFaces. The MetFaces dataset (Karras et al., 2020a) comprises 1,336 high-quality facial images from the Metropolitan Museum of Art's collection (https://metmuseum.github.io/).

BreCaHAD. The BreCaHAD dataset (Aksac et al., 2019), designed for breast cancer histopathology studies, contains 162 images at 1360×1024 resolution. We restructure these into 1,944 partially overlapping crops of 512x512 pixels for our experiments.

FFHQ-2.5K. Using the CLIP visual encoder (Radford et al., 2021), we extract features from the FFHQ images and cluster them into 14 groups using K-means. We select the smallest cluster, which contains 2,500 images (i.e., FFHQ-2.5K), for our study.

C.2. Evaluation Metrics

We adopt the following metrics to evaluate the effectiveness of our method and compare it with the baselines. **Inception Score (IS)** (Salimans et al., 2016) evaluates the quality and diversity of the generated images. It is a common metric in the early stage of GAN development and estimates the quality of the generated images based on the conditional entropy of the class labels predicted by an Inception network. A higher IS score indicates better performance. **Fréchet Inception Distance (FID)** (Heusel et al., 2017) is another widely used metric for measuring the quality of the generated images. FID fits a Gaussian distribution to the hidden activations for each distribution (generated and ground truth) and then computes the Fréchet distance, also known as the Wasserstein-2 distance, between those Gaussians. **Kernel Inception Distance (KID)** (Binkowski et al., 2018) is a metric similar to FID, which is the squared maximum mean discrepancy (MMD) between Inception representations. Unlike FID, it does not assume that the activation distribution has a parametric form, and it has a simple unbiased estimator. It is more informative in situations where the ground truth data is small scale. A lower KID score indicates better performance.

C.3. Setups

We train SNGAN and StyleGAN2 using implementations from Miyato et al. (2018) and Karras et al. (2020b), respectively, and implement a 64×64 version of DCGAN following Radford et al. (2016). All experiments use the same hyperparameter settings, and model performance is evaluated using the implementation from Karras et al. (2020b). For integrating MS^{3}D with other methods, we set λ to 10 in StyleGAN2, while it is set to 100 in other methods due to StyleGAN2's additional constraint terms. We note a discrepancy between reported scores in the literature and our results using the provided code, possibly due to hardware variations or differences between runs.

D. More Qualitative Results

In this section, we present additional qualitative results obtained from OxfordDog, FFHQ-2.5K, MetFaces, and BreCaHAD datasets. We compare the outcomes of StyleGAN2 and StyleGAN2 with MS^{3}D on these datasets, as illustrated in Figs. 18 and 19.

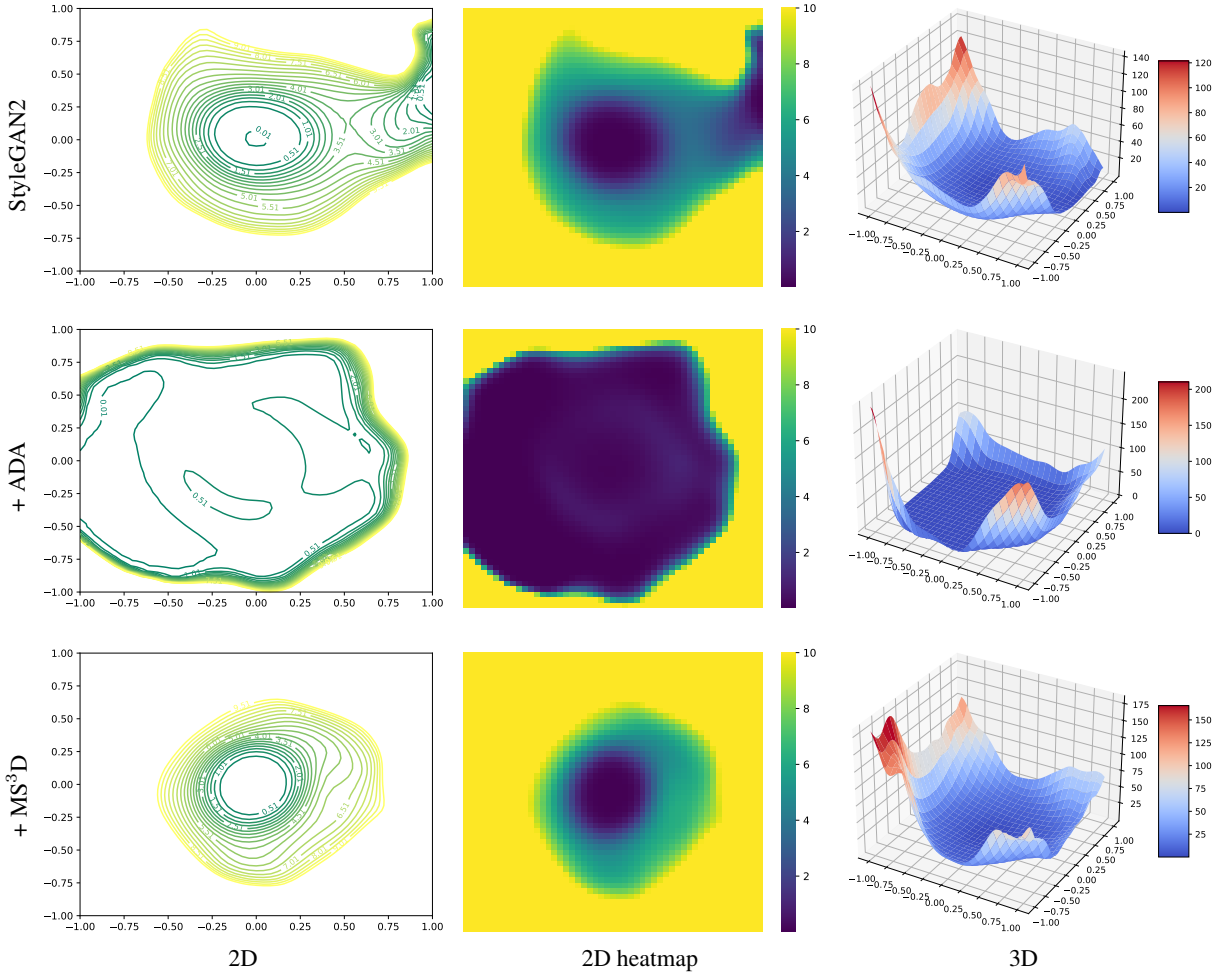


Figure 14. Visualization of the loss landscapes.

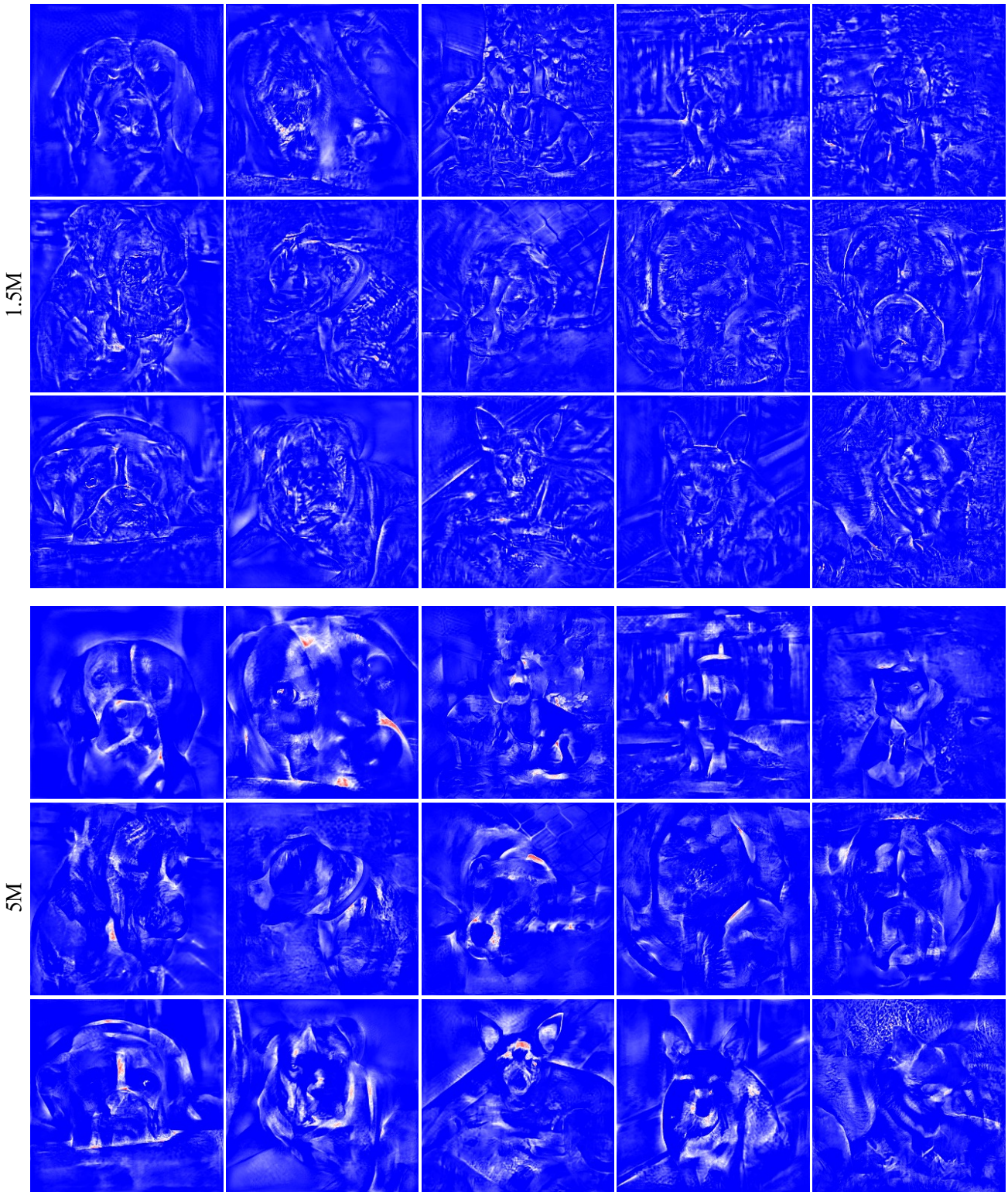


Figure 15. Visualizing $\nabla_x(f(x;\phi))$ on StyleGAN2 (uncurated).

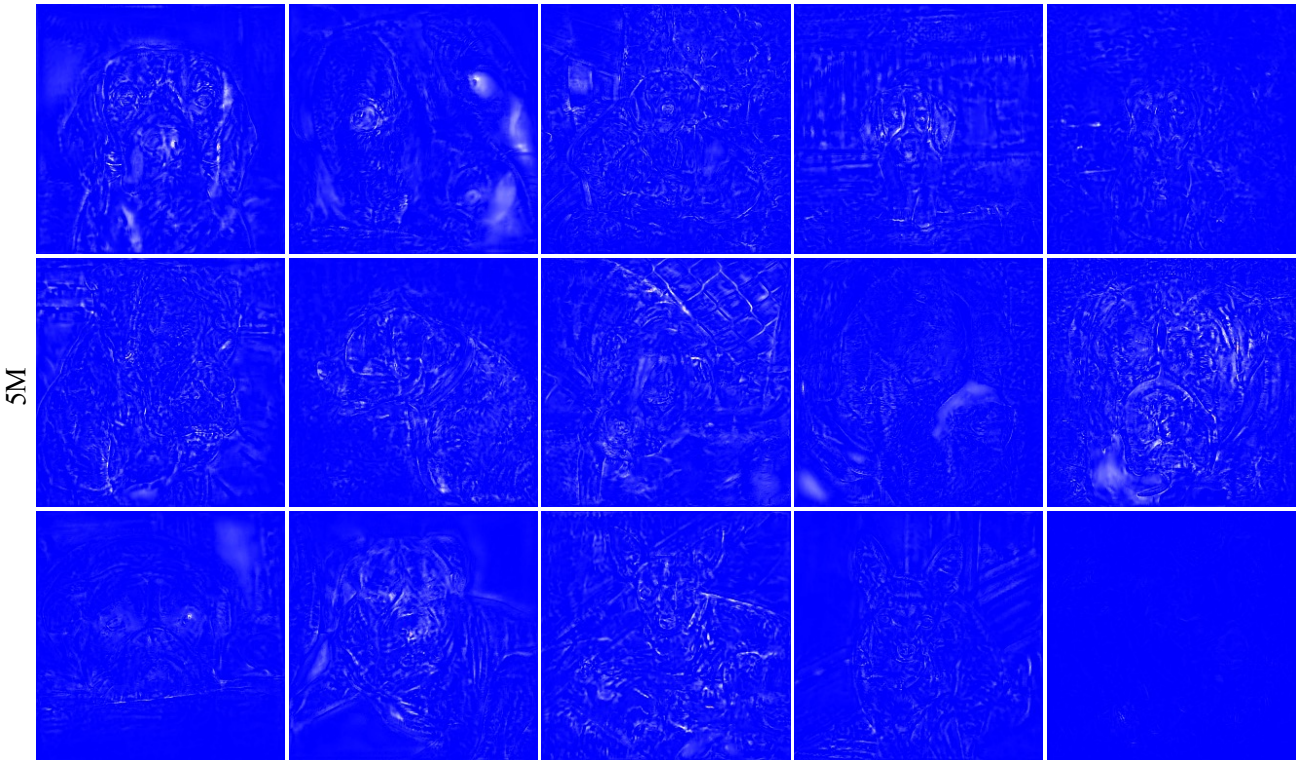


Figure 16. Visualizing $\nabla_x(f(x;\phi))$ on StyleGAN2 + MS^{3}D (uncurated).

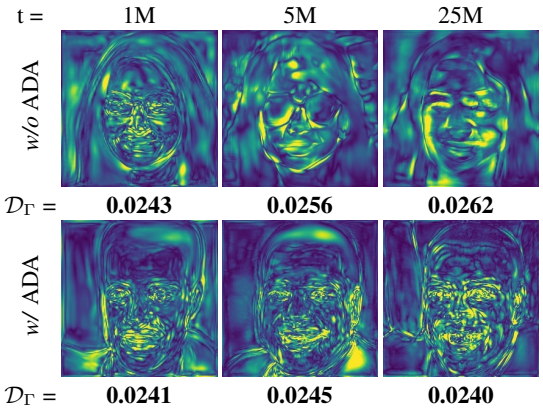


Figure 17. MS^{3}D analysis on natural patterns. The computation of the MS^{3}D descriptor, $\mathcal{D}_\Gamma$, is performed directly on the images obtained from (Karras et al., 2020a)

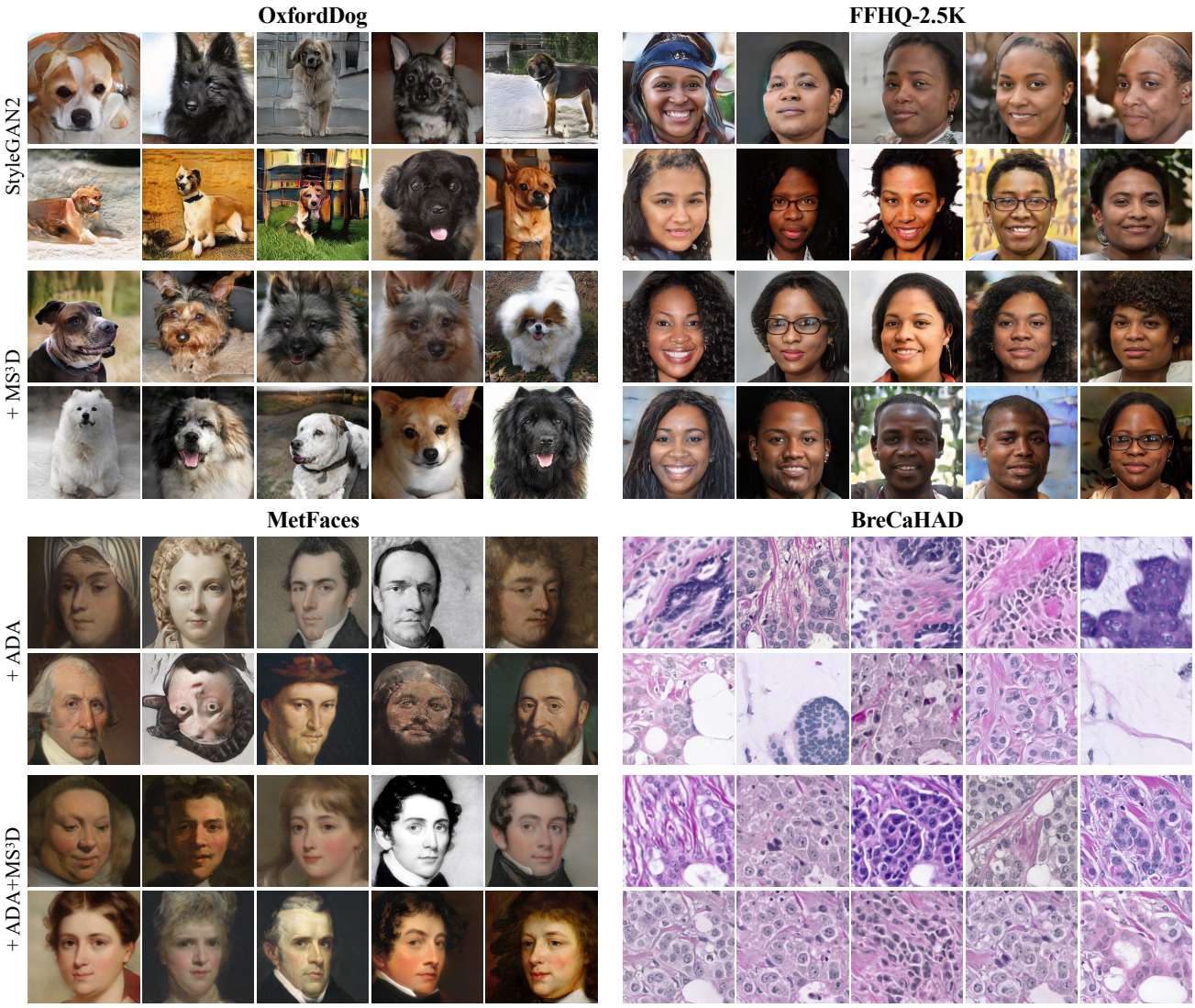


Figure 18. More generated images (best FID).

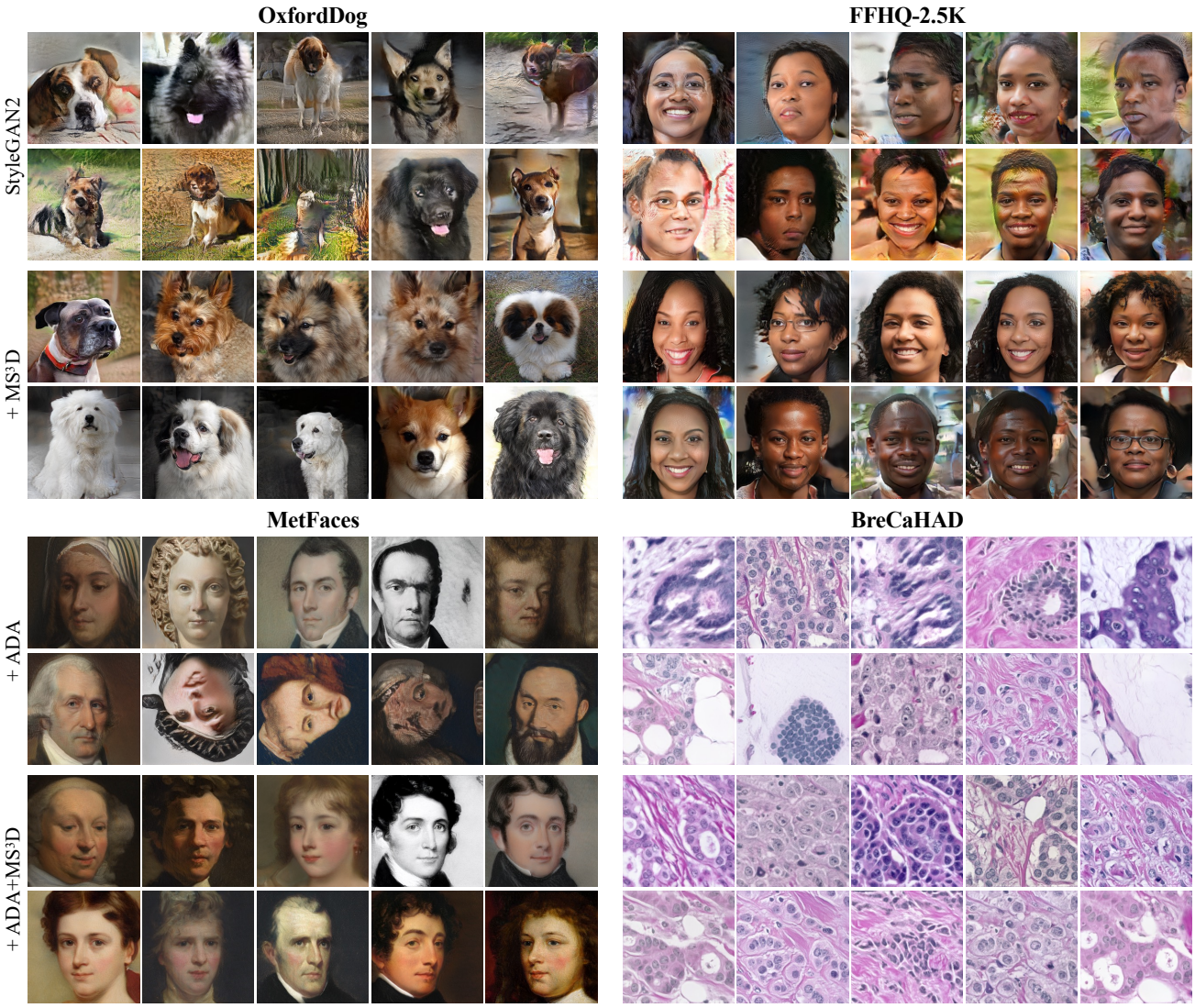


Figure 19. More generated images (last iteration).