Combating Mode Collapse via Offline Manifold Entropy Estimation

Haozhe Liu[1,2], Bing Li[1*], Haoqian Wu[3,4], Hanbang Liang[4], Yawen Huang[2],
Yuexiang Li [2*], Bernard Ghanem[1], Yefeng Zheng[2]

[1] AI Initiative, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia
[2] Jarvis Lab, Tencent, Shenzhen 518057, China
[3] YouTu Lab, Tencent, Shenzhen 518057, China
[4] Shenzhen University, Shenzhen 518060, China
{haozhe.liu;bing.li;bernard.ghanem}@kaust.edu.sa; lianghanbang2019@email.szu.edu.cn;
{linuswu;vicyxli;yawenhuang;yefengzheng}@tencent.com

Abstract

Generative Adversarial Networks (GANs) have shown compelling results in various tasks and applications in recent years. However, mode collapse remains a critical problem in GANs. In this paper, we propose a novel training pipeline to address the mode collapse issue of GANs. Different from existing methods, we propose to generalize the discriminator as feature embedding and maximize the entropy of distributions in the embedding space learned by the discriminator. Specifically, two regularization terms, *i.e.*, Deep Local Linear Embedding (DLLE) and Deep Isometric feature Mapping (DIsoMap), are introduced to encourage the discriminator to learn the structural information embedded in the data, such that the embedding space learned by the discriminator can be well-formed. Based on the well-learned embedding space supported by the discriminator, a non-parametric entropy estimator is designed to efficiently maximize the entropy of embedding vectors, playing as an approximation of maximizing the entropy of the generated distribution. By improving the discriminator and maximizing the distance of the most similar samples in the embedding space, our pipeline effectively reduces the mode collapse without sacrificing the quality of generated samples. Extensive experimental results show the effectiveness of our method which outperforms the GAN baseline, MaF-GAN on CelebA (9.13 *vs.* 12.43 in FID) and surpasses the recent state-of-the-art energy-based model on the ANIMEFACE dataset (2.80 *vs.* 2.26 in Inception score).

Introduction

Generative Adversarial Networks (GANs) have attracted extensive attention in recent years (Schmidhuber 1990, 1991, 2020; Goodfellow et al. 2020). Generally speaking, a GAN consists of a generator network and a discriminator network, where the generator generates samples to fool the discriminator, and the discriminator is trained to discriminate real and generated samples. With such adversarial learning, GANs have shown high-fidelity results in various tasks such as image inpainting (Yu et al. 2018) and photo super-resolution (Li et al. 2019). Nevertheless, GANs suffer from

This work was done when Haozhe Liu interned at Jarvis Lab.
Corresponding Authors: Bing Li and Yuexiang Li.
Code is available at https://github.com/HaozheLiu-ST/MEE.

mode collapse (or training instability) (Mangalam and Garg 2021; Saxena and Cao 2021), hindering their further development in the generative learning community and potential applications.

To alleviate the mode collapse in GANs, many efforts have been devoted to introducing prior knowledge or adding noise on the generator side. For example, the conditional GANs (Mirza and Osindero 2014; Odena, Olah, and Shlens 2017; Brock, Donahue, and Simonyan 2019) handle the mode collapse via introducing class-level prior knowledge into GANs. By splitting the generated distribution into several sub-distributions, the complexity of image synthesis would be further reduced. As the distributions of different classes are quite different, mode collapse can be mitigated to some extent. However, annotating the image classes is time-consuming and deviates from the unsupervised setting of GAN. Compared to conditional information, adding noise into networks is a more flexible solution against mode collapse. Specifically, Style-GANs (Karras, Laine, and Aila 2019; Karras et al. 2020b) add Gaussian noise to the output of each convolutional layer to increase the variance of the generated samples. Despite the rapid progress of GAN-based approaches, mode collapse remains an unsolved and challenging problem in GANs.

Different from GANs, another kind of generative model, namely Energy-Based Models (EBMs) (Grathwohl et al. 2021; Geng et al. 2021; LeCun et al. 2006; Xie et al. 2016; Xie, Zhu, and Nian Wu 2017; Xie, Zhu, and Wu 2019; Xie et al. 2018b), have shown remarkable performances in circumventing the mode collapse. These methods have attracted increasing attention in recent years. Typically, EBMs estimate the density of the target data distribution and are trained via maximum likelihood. The partition function has to be estimated during the training of EBMs by adopting expensive Langevin dynamics, while the partition function is generally intractable. Consequently, since training EBMs is challenging, EMBs suffer from high computational complexity and cannot generate images with competitive fidelity.

In this paper, inspired by the success of EBMs in avoiding the mode collapse, we propose a novel pipeline named MaEM-GANs (manifold entropy maximization) for training GANs through bridging Wasserstein GANs (WGANs) and EBMs. By analyzing the connection between WGANs and EBMs, we discover that mode collapse in GANs can
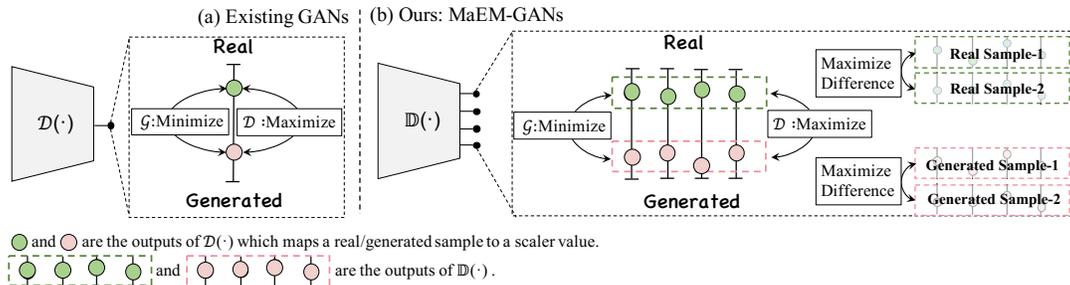
Figure 1: Illustration of our main idea for combating the mode collapse. Different from the existing GANs (a), our MaEM-GANs (b) generalize the discriminator such that it embeds an image into an m-dimensional space, instead of just outputting a scalar value. With the embedding vectors of real/generated samples, we maximize the entropy of distributions in the embedding space learned by our discriminator to prevent the mode collapse.

be avoided by maximizing the entropy of the generated distribution. However, it is nontrivial to directly estimate the entropy of the generated distribution for GANs, especially for large-scale and high-dimensional data. Instead, since the discriminator heavily affects the training quality/stability of GANs (Karras et al. 2020a), we propose to address the entropy maximization on the discriminator side. In particular, we propose to generalize the discriminator to feature embedding, such that it embeds images into a lower-dimensional embedding space (see Fig. 1), different from typical GANs. We then maximize the entropy of distributions in the embedding space learned by the discriminator. To further optimize such surrogate objective in an efficient and simple manner, we propose a module named RB-MaEM based on non-parametric entropy estimator using replay buffer. In addition, we introduce two regularization terms, namely, Deep Local Linear Embedding (DLLE) and Deep Isometric feature Mapping (DIsoMap), which encourages the discriminator to learn the structural information embedded in the data such that the embedding space is well formed. Benefiting from DLLE and DIsoMap, our method, namely MaEM-GAN, maximizes the entropy in the well-learned embedding space to combat the mode collapse in GANs. Experimental results show that the proposed MaEM-GAN outperforms the recent advanced GAN method MaF-GAN (Liu et al. 2021) on CelebA (9.13 *vs.* 12.43 in FID) and surpasses the recent state-of-the-art EBM (Geng et al. 2021) on the ANIMEFACE dataset (2.80 *vs.* 2.26 in Inception score).

Our contributions are summarized as follows:

- We propose a novel training pipeline to address the mode collapse issue in GANs, which effectively alleviates mode collapse without sacrificing the image quality of generated images.

- We show that the mode collapse in GANs can be reduced by generalizing the discriminator as feature embedding and maximizing the entropy of distributions in the embedding space learned by the discriminator.

- Extensive experiments show that our method achieves superior performances in terms of diversity and image quality on various image generation tasks, compared with both state-of-the-art GANs and EBMs.

Related Works

Generative Adversarial Networks. To overcome the mode collapse, numerous methods have been proposed, such as introducing class-level information (Mirza and Osindero 2014; Odena, Olah, and Shlens 2017; Brock, Donahue, and Simonyan 2019; An et al. 2019) or adding noises to different layers (Karras, Laine, and Aila 2019; Karras et al. 2020b). However, these methods need additional annotation, which deviates from the unsupervised setting of GAN, or introduces complex network architecture. Differently, WGANs (Gulrajani et al. 2017; Arjovsky, Chintala, and Bottou 2017) re-designed the learning objective, where the discriminator performs regression, rather than classification. By dynamically modeling the distance between the generated and real distributions, WGANs can reduce the risk for local minima and mitigate the mode collapse to a certain extent. More recently, Realness GANs (Xiangli et al. 2020) and Manifold-preserved GANs (Liu et al. 2021) generalize GANs into a high-dimensional form by mapping the output of the discriminator into a vector to enhance the realness. Since the fidelity of the given image can be judged from different views like ensemble learning, the generator can synthesize images with different attributes to fit the discriminator. The above works indicate that the discriminator is critical for GANs and can be a potential target to combat mode collapse. Motivated by these progresses (AZhang et al. 2020; Xiangli et al. 2020; Liu et al. 2021) with regard to the discriminator, we propose a new pipeline, focusing on designing the simple yet effective constraints on the discriminator side to prevent the mode collapse. Different from the previous methods, we explore how to maximize the entropy of distribution in the embedding space supported by the discriminator, and how to well learn the embedding space. The empirical and theoretical study pointed out that the learning objective of GANs might neglect an intractable entropy term for maximum likelihood, which plays a key point in the mode collapse and can be effectively tackled by high-dimensional GANs. Manifold-preserved GAN (Liu et al. 2021) is the most similar work to ours; however, it neglects the entropy term.

Energy-based Models. There has been a rich history for EBMs, which can be traced back to Hopfield networks

(Hopfield 1982) and Boltzmann machines (Hinton and Sejnowski 1983). However, learning an EBM is difficult, since the partition function (*a.k.a.* normalization constant) is intractable and hard to estimate (Geng et al. 2021). A common solution is based on expensive Markov Chain Monte Carlo (MCMC) sampling by generally adopting Langevin dynamics (Du and Mordatch 2019; Xie et al. 2018a) and Gibbs sampling (Carlo 2004; De Sa, Chen, and Wong 2018) to estimate the partition function. Some strict requirements, including parameter tuning (Grathwohl et al. 2020), early stopping of MCMC (Nijkamp et al. 2019), and avoiding the use of modern deep modules (*e.g.*, self-attention, dropout and batch/layer normalization) (Grathwohl et al. 2020), are adopted to mitigate the training instability issue caused by MCMC. These hard requirements limit the capacity of the deep model and might reduce its applicability to some large-scale datasets. More recently, an MCMC-free EBM training strategy (Grathwohl et al. 2021) was proposed, where a generator is employed as a sampler to achieve amortized training of EBM. In particular, the output of the discriminator (also regarded as EBM) is a single scalar, hence, it is hard to directly estimate the entropy, resulting in complicated processing based on variational inference (Grathwohl et al. 2021) or Jacobi-determinant (Geng et al. 2021). Inspired by the amortized training strategy, we generalize WGAN into a manifold representation and adopt a replay buffer strategy to directly estimate the entropy of the manifold representation. Compared with the state-of-the-art EBMs, the proposed method is simple but effective to estimate entropy, leading to stronger diversity and fidelity in image generation.

Method

The purpose of our method is to combat the mode collapse in GANs. To this end, we first bridge WGANs and EBMs, since EBMs have shown remarkable performance in avoiding the mode collapse. By analyzing the connections between WGANs and EBMs, we discover that mode collapse in GANs can be prevented via maximizing the entropy of the generated distribution. However, it is nontrivial to directly estimate the entropy of the generated distribution. Instead, we propose a new training pipeline MaEM-GANs to approximate this objective. In particular, we generalize the discriminator to feature embedding, to embed images into a low-dimensional space. With the generalized discriminator, we propose a surrogate objective to maximize the entropy of the distribution in the embedding space learned by the discriminator. RB-MaEM is proposed to efficiently optimize the surrogate objective. Furthermore, we introduce two regularization terms, *i.e.*, DLLE and DIsoMap, to ensure that the embedding space is well-formed and captures the underlying manifold embedded in the high-dimensional data.

Problem Definition

Based on (Grathwohl et al. 2021; Geng et al. 2021), we revisit EBMs and WGANs. By analyzing the connections between them, we discover that the mode collapse can be prevented by maximizing the entropy of generated distribution.
Definition 1. An EBM can be represented from Gibbs density:

$$p_\theta(x) = \frac{e^{f_\theta(x)}}{Z(\theta)}, \tag{1}$$

where $f_\theta : \mathbb{R}^{h \times w} \to \mathbb{R}^1$ (h and w denote the height and width of the image, respectively) and $Z(\theta) = \int e^{f_\theta(x)} dx$ ($Z(\theta)$ is the partition function or normalizing constant). EBMs are trained by maximum likelihood estimation:

$$\mathcal{L}_{ebm}(\theta) := -\mathbb{E}_{x \sim \mathcal{P}_r}[\log(\frac{e^{f_\theta(x)}}{Z(\theta)})] = -\mathbb{E}_{x \sim \mathcal{P}_r}[f_\theta(x)] + \log Z(\theta), \tag{2}$$

where $\mathcal{P}_r$ is the data distribution.
Definition 2. The loss function of WGANs can be defined as

$$\mathcal{L}_{wgan}(\theta) := \mathbb{E}_{x \sim \mathcal{P}_g}[\mathcal{D}(x)] - \mathbb{E}_{x \sim \mathcal{P}_r}[\mathcal{D}(x)], \tag{3}$$

where $\mathcal{P}_g$ is the generated distribution. Concretely, the discriminator $\mathcal{D}(\cdot)$ is trained by minimizing $\mathcal{L}_{wgan}(\theta)$, while the generator $\mathcal{G}(\cdot)$ is driven to maximize $\mathcal{L}_{wgan}(\theta)$.
Proposition 1. $\mathcal{L}_{wgan}(\theta)$ acts as a lower bound of $\mathcal{L}_{ebm}(\theta)$ by maximizing the entropy of the generated distribution $\mathcal{H}(\mathcal{P}_g)$.
Proof: Given the probability density function $q(x)$ of $\mathcal{P}_g$ as input, an inequality can be derived as

$$\log Z(\theta) \geq \log Z(\theta) - \mathrm{KL}(q(x)\|p_\theta(x)), \tag{4}$$

where $\mathrm{KL}(\cdot\|\cdot)$ denotes the KL-divergence. With respect to $p_\theta(x) = e^{f_\theta(x)}/Z(\theta)$, $\log Z(\theta)$ can be re-written as

$$\begin{aligned}
\log Z(\theta) &\geq \log Z(\theta) + \int_x q(x) \log(\frac{e^{f_\theta(x)}}{Z(\theta)q(x)}) \\
&= \mathbb{E}_{x \sim \mathcal{P}_g}[f_\theta(x)] + H(\mathcal{P}_g),
\end{aligned} \tag{5}$$

where $\mathcal{H}(\mathcal{P}_g)$ is the entropy of $\mathcal{P}_g$. Note that model $f_\theta(x)$ in EBMs is equivalent to $\mathcal{D}(\cdot)$ in GANs. Hence, the connection between $\mathcal{L}_{wgan}$ and $\mathcal{L}_{ebm}$ can be established as

$$\begin{aligned}
\mathcal{L}_{ebm}(\theta) &= -\mathbb{E}_{x \sim \mathcal{P}_r}[f_\theta(x)] + \log Z(\theta) \\
&\geq \mathcal{L}_{wgan}(\theta) + \mathcal{H}(\mathcal{P}_g)
\end{aligned} \tag{6}$$

Based on Eq. (6), $\mathcal{L}_{ebm}(\theta)$ is the upper bound of $\mathcal{L}_{wgan}(\theta)$ and an entropy term $\mathcal{H}(\mathcal{P}_g)$. Generally, such an upper bound is very tight, because of the distance between $\mathcal{L}_{wgan}(\theta) + \mathcal{H}(\mathcal{P}_g)$ and $\mathcal{L}_{ebm}(\theta)$ is the KL-divergence term $\mathrm{KL}(q(x)\|p_\theta(x))$ (Grathwohl et al. 2021) (see Eq. (4)). That is, for the generator, we can maximize $\mathcal{L}_{wgan}$ and the entropy term $\mathcal{H}(\mathcal{P}_g)$ to approximate L_{ebm} that well alleviates the mode collapse issue. Intuitively, $\mathcal{H}(\mathcal{P}_g)$ plays a key role to tackle the mode collapse problem, which is generally neglected by existing GANs.

Manifold Representation for WGANs

As discussed above, mode collapse can be prevented by maximizing the entropy of the generated distribution. However, directly estimating the entropy $\mathcal{H}(\mathcal{P}_g)$ of the generated distribution is intractable, due to the large quantity and
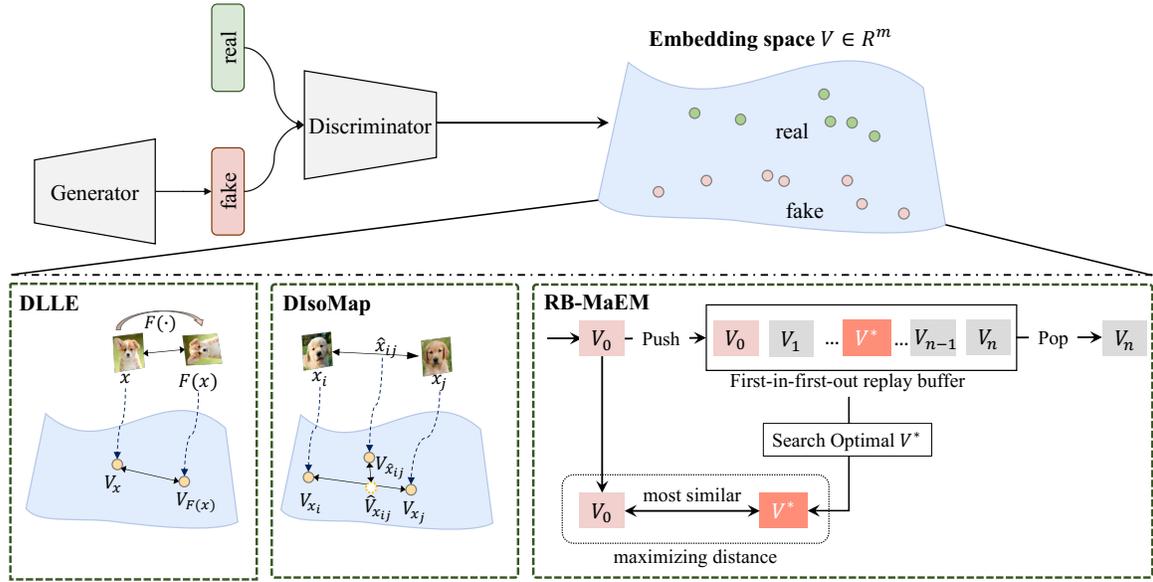
Figure 2: The pipeline of the proposed MaEM-GANs. Our discriminator embeds an input into an m-dimensional code V, instead of a scalar value. To preserve the structural information of manifolds embedded in the input data, we introduce two regularization terms DLLE and DIsoMap. Within the embedding space learned by the discriminator, RB-MaEM module is proposed to maximize the entropy of distributions by maximizing the distance of most similar samples in the replay buffer.

high dimensionality of data samples. We address the above problem on the discriminator side, motivated by the fact the discriminator heavily affects the training stability of GANs (Karras et al. 2020a). Specifically, we propose to generalize the discriminator from the perspective of the manifold and maximize the entropy of distributions in the embedding space supported by the discriminator.

Typical GANs (Arjovsky, Chintala, and Bottou 2017; Goodfellow et al. 2014; Gulrajani et al. 2017; Li et al. 2021) treat the discriminator as a classifier, which classifies an image x as real or fake according to a *scalar* value $\mathcal{D}(x)$. Different from the scalar-based discriminators, our generalized discriminator $\mathbb{D}(\cdot)$ learns a mapping which transforms an input image to an m-dimensional embedding space: $\mathbb{D}(\cdot)$: $\mathbf{R}^{h \times w} \to \mathbf{R}^m$, where h and w are the height and weight of an image x, respectively, and $m \geq 1$.

The embedding space $\mathbb{D}(\cdot)$ is expected to capture informative characteristics of images. In other words, each dimension of the embedding space $\mathbb{D}(\cdot)$ corresponds to a critical attribute of images, such as color, texture, and structure. Thus, compared with the scalar-based discriminators, our generalized discriminator $\mathbb{D}$ can provide a more comprehensive representation for the data.

With our generalized discriminator $\mathbb{D}$, we reformulate the learning objective of WGANs in Eq. (3). Although the output of our discriminator $\mathbb{D}$ is a vector, our discriminator determines whether an image x is real/fake according to the average value of $\mathbb{D}(x)$. The new objective function $\mathcal{L}_{MaF}$

can be formulated as:

$$\mathcal{L}_{MaF} \triangleq \mathop{\mathbb{E}}_{z \sim \mathcal{P}_z}\left[\frac{1}{m}\sum_k \mathbb{D}_k(\mathcal{G}(z))\right] - \mathop{\mathbb{E}}_{x \sim \mathcal{P}_r}\left[\frac{1}{m}\sum_k \mathbb{D}_k(x)\right],$$
(7)

where the standard normal distribution $\mathcal{N}(0,1)$ is used for p_z and $\frac{1}{m}\sum_k \mathbb{D}_k(x)$ indicates the realness of x.

Replay-Buffer-based Manifold Entropy Estimation

With our generalized discriminator $\mathbb{D}(\cdot)$, we design a surrogate objective to maximize the entropy of distributions in the embedding space supported by $\mathbb{D}(\cdot)$.

Given the ith image x_i, we represent it as an embedding code $V_i = \mathbb{D}(x_i)$ using our discriminator. With such representation, we aim to maximize the entropy of the embedding codes $\{V_i\}$'s distribution. Inspired by non-parametric entropy estimator (Beirlant et al. 1997), our insight is that if the distance between neighboring samples is maximized in a manifold, the data points in such a manifold will follow a uniform distribution, which maximizes the entropy of data distributions. We hence propose RB-MaEM module to efficiently conduct this objective using a replay buffer. As shown in Fig. 2, RB-MaEM employs a first-in-first-out replay buffer $\mathcal{R}$ to store embedded codes $\{V_0, ..., V_n\}$ of $n+1$ sample images, where V_0 is the code of the 0th image lying at the head of $\mathcal{R}$ and V_n is the last code at the tail of $\mathcal{R}$. We then search for an embedding code V^* which is the most similar to V_0 from $V_1, ..., V_n$ in buffer $\mathcal{R}$. With V^*, we maximize the entropy of the distribution by maximizing the distance of V_0 and its most similar code V^*:

$$\mathcal{L}_{ent} = \frac{V_0 \cdot V^*}{||V_0|| \cdot ||V^*||} + \lambda||V_0||,$$
(8)

where $V^* = \arg\max_{V_i \in \mathcal{R}} (\frac{V_0 \cdot V_i}{||V_0|| \, ||V_i||})$ obtained based on cosine similarity, and $\lambda ||V_0||$ is a regularization term to ensure the stabilization of the discriminator.

We further show that maximizing the entropy of the generated distributions in the embedding space supported by $\mathbb{D}(\cdot)$ approximately maximizes the entropy $\mathcal{H}(\mathcal{P}_g)$ of generated distribution which is stated in **Proposition** 1.

Proposition 2. The entropy of the distribution $\mathcal{P}_g$ in the embedding space learned by the discriminator $\mathbb{D}(\cdot)$ is the lower bound of $\mathcal{H}(\mathcal{P}_g)$, based on the non-parametric entropy estimator and Lipschitz continuity.

Proof: Based on the non-parametric entropy estimator (Beirlant et al. 1997), $\mathcal{H}(\mathcal{P}_g)$ can be maximized by

$$\max_{x_i \sim \mathcal{P}_g} \mathbb{E} [\log(d_\mathcal{X}(x_i, x_i^*))], \qquad (9)$$

where $x_i^* \triangleq \arg\min_{x_j \sim \mathcal{P}_g} d_\mathcal{X}(x_i, x_j)$, and $d_\mathcal{X}$ measures the distance between samples in image space $\mathcal{X}$. Based on the convergence of WGAN, $\mathbb{D}$ is under Lipschitz continuity:

$$K d_\mathcal{X}(x_i, x_j) \geq d_\mathbb{D}(\mathbb{D}(x_i), \mathbb{D}(x_j)), \qquad (10)$$

where K is the Lipschitz constant, and $d_\mathbb{D}$ is the metric in the embedding space supported by $\mathbb{D}$. Since K is typically set to be 1, $d_\mathbb{D}(\mathbb{D}(x_i), \mathbb{D}(x_j))$ in the space of the discriminator can be regarded as the lower bound of the distance between x_i and x_i^* in image space. By using $d_\mathbb{D}(\mathbb{D}(x_i), \mathbb{D}(x_j))$ (*i.e.*, lower bound of $d_\mathcal{X}(x_i, x_j)$) in Eq. 9 , we have

$$\max_{x_i \sim \mathcal{P}_g} \mathbb{E} [\log(d_\mathcal{X}(x_i, x_i^*))] \geq$$
$$\max_{x_i \sim \mathcal{P}_g} \mathbb{E} [\log(d_\mathbb{D}(\mathbb{D}(x_i), \mathbb{D}(x_i^*)))] \qquad (11)$$

Therefore, the entropy of generated distribution in the image space can be approximately maximized by maximizing the entropy in the embedding space.

Manifold Regularization

The performance of the proposed RB-MaEM depends on the quality of the embedding space learned by $\mathbb{D}(\cdot)$. To ensure that the embedding space captures underlying manifolds embedded in the high-dimensional data, we introduce DLLE and DIsoMap to regularize the learning of $\mathbb{D}$.

DLLE. The first regularization term DLLE enforces the discriminator $\mathbb{D}$ to preserve the nonlinear structure of high-dimensional data by using the local symmetries of linear reconstructions. In particular, DLLE is established upon a simple geometric intuition (Roweis and Saul 2000), *i.e.*, a data sample and its neighbors lie on or are close to a locally linear region of the manifold learned by $\mathbb{D}(\cdot)$. Inspired by representation learning (Chen et al. 2020; AZhang et al. 2020), we characterize a different view of x as the neighbor of x. Then, the learning objective $\mathcal{L}_{LLE}$ of DLLE enforces x and its neighbor to be similar in the embedding space of $\mathbb{D}(\cdot)$:

$$\mathcal{L}_{LLE} \triangleq \mathbb{E}_{x \sim \mathcal{P}_r, \mathcal{P}_g} [||\mathbb{D}(\mathcal{F}(x)) - \mathbb{D}(x)||_2] \qquad (12)$$

$$= \mathbb{E}_{x \sim \mathcal{P}_r, \mathcal{P}_g} [||V_{F(x)} - V_x||_2], \qquad (13)$$

where $\mathcal{F}(\cdot)$ is the image transformation, such as rotation, adding Gaussian noise, and adversarial noise. By capturing the invariance between x and its neighbors, our DLLE helps the embedding space of $\mathbb{D}(\cdot)$ preserve the local geometry in the original data, leading to meaningful representations.

DIsoMap. Inspired by MaF-GANs (Liu et al. 2021), we introduce DIsoMap to improve the embeddings' quality. In DIsoMap, we preserve the relationship between the embedding combination of $\{\mathbb{D}(x_i), \mathbb{D}(x_j)\}$ and the embedding of the combination of $\{x_i, x_j\}$ with cosine similarity. Our DIsoMap further preserves the topological structure of different samples for manifold learning.

Experimental Results and Analysis

We implement our MaEM-GAN using the public PyTorch toolbox on eight NVIDIA V100 GPUs. To evaluate the performance of the proposed method, extensive experiments are carried on four publicly available datasets with different image sizes, including **CIFAR-10** (32×32 pixels) (Krizhevsky, Hinton et al. 2009),[1] **ANIMEFACE** (64×64),[2] **CelebA** (256×256) (Liu et al. 2015),[3] and **FFHQ** (1024×1024) (Karras, Laine, and Aila 2019).[4] All the experimental settings, such as optimizer, network architecture and learning rate, are identical to the public benchmarks (Geng et al. 2021; Liu et al. 2021; Xiangli et al. 2020; Karras, Laine, and Aila 2019). Detailed information on the implementation of our MaEM-GAN can be found in *Supplementary Materials*.

Evaluation metrics. We use Fréchet Inception Distance (FID) (Heusel et al. 2017) and Inception Score (IS) (Salimans et al. 2016) to evaluate the quality of generated images. Following state-of-the-art approaches (Geng et al. 2021; Xiangli et al. 2020), we use FID as our main evaluation metric.

To evaluate the effectiveness of our method on alleviating the mode collapse of GANs, we introduce F_8 and a new metric named *I-Variance*, which measures the diversity of generated images to represent the degree of mode collapse. In other words, the larger diversity of the generated images indicates a lower degree of mode collapse. Hence, I-Variance is defined as the standard deviation of generated distributions, where a generated image is represented by the extracted feature using Inception-V3 (Szegedy et al. 2016):

$$\text{I-Variance} \triangleq \sqrt{\mathbb{E}_{x \in \mathcal{P}_g} \big[||\mathcal{T}(x) - \mathbb{E}_{x \in \mathcal{P}_g} [\mathcal{T}(x) ||_2 \big]}, \qquad (14)$$

where $\mathcal{T}(\cdot)$ is the Inception-V3 pre-trained on ImageNet. In all experiments, 50,000 images are randomly sampled to calculate FID, IS, and I-Variance.

Ablation Studies and Analysis

Ablation studies. To conduct ablation studies, we remove DIsoMap, DLLE, and RB-MaEM from our method to establish a baseline, *i.e.*, a high-dimensional WGAN with $\mathcal{L}_{MaF}$.

---

[1] https://www.cs.toronto.edu/ kriz/cifar.html

[2] https://www.kaggle.com/splcher/animefacedataset

[3] https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html

[4] https://github.com/NVlabs/ffhq-dataset

Baseline	DIsoMap	DLLE	RB-MaEM	FID↓
✓	✗	✗	✗	>100
✓	✓	✗	✗	51.96
✓	✓	✓	✗	31.70
✓	✓	✗	✓	30.73
✓	✓	✓	✓	**29.22**

Table 1: The ablation study of the proposed method on CIFAR-10 in terms of FID.

w/ DLLE	Adv. Noise	Rot. and Gau. Noise	FID ↓
✓	✗	✗	30.73
✓	✓	✗	33.04
✓	✗	✓	**29.22**

Table 2: FID of the proposed method with different image transformation strategies $\mathcal{F}(\cdot)$ of DLLE on CIFAR-10. (Adv.–Adversarial; Rot.–Rotation; Gau.–Gaussian)

Without our proposed components, the performance of this baseline is unsatisfactory in terms of FID score, as listed in Table 1. More specifically, the discriminator and generator of the baseline with high dimension are trained in an imbalanced mode, which makes the training process unstable. In contrast, the performance of our method is significantly improved by adding each proposed component to the baseline, validating the effectiveness of our components. As shown in Table 1, DIsoMap is the key component to stabilize the training process, since this term can dynamically ensure the distance between the real and generated distributions to be tractable. Hence, DIsoMap drives the baseline to achieve an FID of 51.96. However, this term might lead to a trivial solution, *i.e.*, each dimension outputs the same value. To overcome the issue, DLLE and RB-MaEM are added and help the discriminator yield non-trivial representation, which further improves our performance to 31.70 and 30.73 in FID, respectively. By incorporating all proposed components, our method achieves the best FID score of 29.22.

Impact of Hyper-parameters. For DLLE, Tab. 2 investigates different strategies of generating the neighbors of a given sample, where the first strategy adds the adversarial perturbation into the image. The second strategy employs image augmentations *i.e.*, rotation and adding Gaussian noises, which is simpler yet achieves better performance in FID than the first one. In addition, Table 3 lists our results using different buffer sizes. Our method will be degraded if the buff size is too small to store enough meaningful samples for maximizing the entropy. When the buffer size is 1024, our method achieves the best performance.

The effectiveness of the discriminator with manifold representation. In our method, the discriminator yields a vector to measure the realness of generated images. Each element in the vector measures the sample-wise realness from a specific attribute. To demonstrate the effectiveness of manifold representation, we feed 50,000 generated images to our discriminator and obtain the corresponding vectors. Fig. 3 shows the 'realest' samples ranked along different
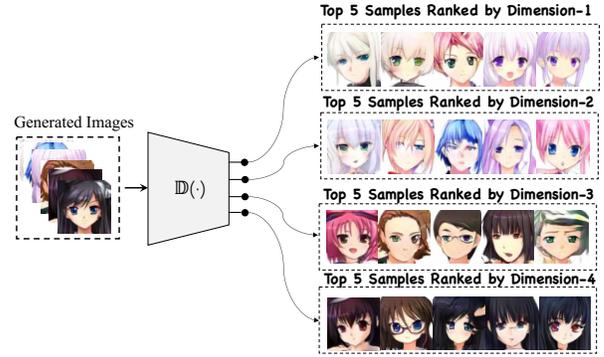
Figure 3: Top five generated images ranked by the individual dimension of the discriminator on ANIMEFACE. In each row, a generated image is ranked according to the cth dimension value of its embedded code V extracted by our discriminator ($c = 1, 2, 3, 4$).

Buffer size	128	512	1024	2048
FID	31.58	32.36	**29.22**	31.87

Table 3: FID ↓ of our method using different replay buffer sizes on CIFAR-10.

elements/dimensions of the obtained vectors. We can observe that the top samples ranked by different dimensions of our discriminator's output exhibit different attributes such as color and style. Hence, introducing manifold representation is effective to estimate the entropy and helps the discriminator to assess the realness from different aspects.

The effectiveness of RB-MaEM on alleviating mode collapse. We verify the importance of RB-MaEM in alleviating the mode collapse issue by measuring the diversity of generated images, where the I-Variance score is adopted to quantitatively assess the diversity of generated images. Table 4 shows that our method without RB-MaEM only achieves an I-Variance score of 3.41. In contrast, the I-Variance score increases to 4.88 by adding RB-MaEM. This shows that our RB-MaEM can significantly improve the diversity of generated images and effectively alleviate mode collapse.

Comparison with State-of-the-art Methods

To show the superiority of our method, we compare our method with recent state-of-the-art GAN models and energy-based models. Note that the results of all baseline methods are duplicated from the existing benchmarks (Xiangli et al. 2020; Geng et al. 2021) without re-implementation.

Comparison with GAN models. Table 5 shows the comparison results on CIFAR-10 and CelebA datasets. We compare our method with the state-of-the-art GANs, including WGAN (Arjovsky, Chintala, and Bottou 2017), Hinge-GAN (Zhao, Mathieu, and LeCun 2017), LSGAN (Mao et al. 2017), Std-GAN (Goodfellow et al. 2014), WGAN-GP (Gulrajani et al. 2017), Realness GAN (Xiangli et al. 2020), and the family of MaF-GANs (Liu et al. 2021). Compared with these methods, our method achieves the best FID score

Configuration	CIFAR-10	ANIMEFACE
Baseline	3.41 ± 0.03	2.46 ± 0.03
+ RB-MaEM	$\mathbf{4.88 \pm 0.04}$	$\mathbf{2.59 \pm 0.04}$

Table 4: The I-Variance score $\uparrow (\times 10^{-3})$ of our method on CIFAR-10 and ANIMEFACE. I-Variance score denotes the variance of the class-wise outputs from Inception-V3.

Model	CIFAR-10	CelebA
WGAN	55.96	-
HingeGAN	42.40	25.57
LSGAN	42.01	30.76
Std-GAN	38,56	27.02
WGAN-GP	41.86	70.28
Realness GAN-Obj.1	36.73	-
Realness GAN-Obj.2	34.59	23.51
Realness GAN-Obj.3	36.21	-
MaF-CwGAN	39.24	-
MaF-DwGAN	33.73	-
MaF-$\mathbb{E}$wGAN	30.85	12.43
Ours	**29.22**	**9.14**

Table 5: FID $\downarrow$ of GAN models on CelebA and CIFAR-10. All methods use the same backbone of DCGAN (Radford, Metz, and Chintala 2016).

Model	FID $\downarrow$	IS $\uparrow$
Unconditional BigGAN	16.04	9.10
Unconditional BigGAN + Ours	**13.86**	**9.27**

Table 6: FID and Inception Score of BigGAN (Brock, Donahue, and Simonyan 2019) and our method on CIFAR-10.

Model	FID $\downarrow$
StyleGAN-V1 (Karras, Laine, and Aila 2019)	4.40
StyleGAN-V2 (Karras et al. 2020b)	2.84
StyleGAN-V2 + Ours	**2.67**

Table 7: FID of StyleGAN-V1/2 and our method on FFHQ.

Model	IS $\uparrow$	FID $\downarrow$	$F_8 \uparrow$
MEG	2.20	9.31	0.95
VERA	2.15	41.00	0.52
EBM-0GP	2.26	20.53	0.89
EBM-BB	2.26	12.75	0.94
DDPM*	2.18	8.81	0.94
WGAN-0GP	2.22	9.76	0.95
Ours	**2.80**	**8.62**	**0.98**

Table 8: FID, IS and F_8 of EBMs, diffusion models and our method on ANIMEFACE.

on both datasets. Specifically, MaF-GAN is a recent work most related to ours, which also generalizes WGAN into a high-dimensional form. Our method surpasses MaF-GAN by a large margin since our method explicitly maximizes the entropy of distribution in the embedding space of the discriminator. Furthermore, we demonstrate that our method facilitates training very deep GAN architectures even with complicated training strategies. In particular, our method is incorporated into StyleGAN-V2 (Karras et al. 2020b) and BigGAN (Brock, Donahue, and Simonyan 2019) by adding the proposed learning objective to their methods. And the consistent improvements can be found in Tables 5, 7 and 6.

Comparison with energy-based models. We compare our method with representative energy-based models, including MEG (Kumar et al. 2019), VERA (Grathwohl et al. 2021), EBM-0GP (Geng et al. 2021) and EBM-BB (Geng et al. 2021). For a more comprehensive study, we also include Denoising Diffusion Probabilistic Model (DDPM) (Ho, Jain, and Abbeel 2020) and WGAN-0GP (Thanh-Tung, Tran, and Venkatesh 2019). DDPM can be regarded as an upper bound for generation performance since this method generates images by expensive iterative optimization. Following (Geng et al. 2021), we conduct the experiments on the AnimeFace dataset. Table 8 shows that our method outperforms all energy-based models with the highest IS (2.80) and F_8 (0.98), and lowest FID (8.62), demonstrating that our method not only ensures the fidelity of the generated images but also significantly improves their diversity.

Figure 4: The generated samples of the proposed method on (a) CIFAR-10 (32×32), (b) ANIMEFACE (64×64), (c) CelebA (256×256) and (d) FFHQ (1024×1024).

Conclusion

In this paper, we proposed a novel method to alleviate mode collapse in GANs. Our method generalizes the discriminator as a feature embedder, and mode collapse in GANs can be alleviated by maximizing the entropy of distributions in the embedding space learned by the discriminator. Two manifold regularization terms were introduced to preserve the information in the manifold embedded in the data. Based on the well-learned embedding space, a replay-buffer-based entropy estimator was proposed to maximize the diversity of samples in the embedding space. By improving the discriminator and maximizing the entropy of distributions in the embedding space, our method effectively reduces the mode collapse without sacrificing the quality of generated samples.

Acknowledgments

References

An, D.; Guo, Y.; Lei, N.; Luo, Z.; Yau, S.-T.; and Gu, X. 2019. Ae-ot: a new generative model based on extended semi-discrete optimal transport. In *ICLR*.

Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, 214–223. PMLR.

AZhang, H.; Zhang, Z.; Odena, A.; and Lee, H. 2020. Consistency regularization for generative adversarial networks. In *ICLR*.

Beirlant, J.; Dudewicz, E. J.; Györfi, L.; Van der Meulen, E. C.; et al. 1997. Nonparametric entropy estimation: An overview. *International Journal of Mathematical and Statistical Sciences*, 6(1): 17–39.

Brock, A.; Donahue, J.; and Simonyan, K. 2019. Large scale GAN training for high fidelity natural image synthesis. *ICLR*.

Carlo, C. M. 2004. Markov Chain Monte Carlo and Gibbs sampling. *Lecture notes for EEB*, 581: 540.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, 1597–1607. PMLR.

De Sa, C.; Chen, V.; and Wong, W. 2018. Minibatch Gibbs sampling on large graphical models. In *International Conference on Machine Learning*, 1173–1181.

Du, Y.; and Mordatch, I. 2019. Implicit generation and modeling with energy based models. *Advances in Neural Information Processing Systems*, 32.

Geng, C.; Wang, J.; Gao, Z.; Frellsen, J.; and Hauberg, S. 2021. Bounds all around: training energy-based models with bidirectional bounds. *Advances in Neural Information Processing Systems*, 34.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative Adversarial Nets. *Advances in Neural Information Processing Systems*, 27.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.

Grathwohl, W.; Wang, K.-C.; Jacobsen, J.-H.; Duvenaud, D.; Norouzi, M.; and Swersky, K. 2020. Your classifier is secretly an energy based model and you should treat it like one. *International Conference on Learning Representations*.

Grathwohl, W. S.; Kelly, J. J.; Hashemi, M.; Norouzi, M.; Swersky, K.; and Duvenaud, D. 2021. No MCMC for me: Amortized sampling for fast and stable training of energy-based models. In *International Conference on Learning Representations*.

Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; and Courville, A. C. 2017. Improved training of Wasserstein GANs. *Advances in Neural Information Processing Systems*, 30.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. *Advances in Neural Information Processing Systems*, 30.

Hinton, G. E.; and Sejnowski, T. J. 1983. Optimal perceptual inference. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, volume 448, 448–453. Citeseer.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33: 6840–6851.

Hopfield, J. J. 1982. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8): 2554–2558.

Karras, T.; Aittala, M.; Hellsten, J.; Laine, S.; Lehtinen, J.; and Aila, T. 2020a. Training generative adversarial networks with limited data. *Advances in Neural Information Processing Systems*, 33: 12104–12114.

Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for Generative Adversarial Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4401–4410.

Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2020b. Analyzing and improving the image quality of StyleGAN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8110–8119.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images. *Handbook of Systemic Autoimmune Diseases*, 1(4).

Kumar, R.; Ozair, S.; Goyal, A.; Courville, A.; and Bengio, Y. 2019. Maximum entropy generators for energy-based models. *arXiv preprint arXiv:1901.08508*.

LeCun, Y.; Chopra, S.; Hadsell, R.; Ranzato, M.; and Huang, F. 2006. A tutorial on energy-based learning. *Predicting structured data*, 1(0).

Li, B.; Zhu, Y.; Wang, Y.; Lin, C.-W.; Ghanem, B.; and Shen, L. 2021. AniGAN: Style-Guided Generative Adversarial Networks for Unsupervised Anime Face Generation. *IEEE Transactions on Multimedia*.

Li, Z.; Yang, J.; Liu, Z.; Yang, X.; Jeon, G.; and Wu, W. 2019. Feedback network for image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3867–3876.

Liu, H.; Liang, H.; Hou, X.; Wu, H.; Liu, F.; and Shen, L. 2021. Manifold-preserved GANs. *arXiv preprint arXiv:2109.08955*.

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

Mangalam, K.; and Garg, R. 2021. Overcoming Mode Collapse with Adaptive Multi Adversarial Training. *arXiv preprint arXiv:2112.14406*.

Mao, X.; Li, Q.; Xie, H.; Lau, R. Y.; Wang, Z.; and Paul Smolley, S. 2017. Least squares generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 2794–2802.

Mirza, M.; and Osindero, S. 2014. Conditional Generative Adversarial Nets. *arXiv preprint arXiv:1411.1784*.

Nijkamp, E.; Hill, M.; Zhu, S.-C.; and Wu, Y. N. 2019. Learning non-convergent non-persistent short-run MCMC toward energy-based model. *Advances in Neural Information Processing Systems*, 32.

Odena, A.; Olah, C.; and Shlens, J. 2017. Conditional image synthesis with auxiliary classifier GANs. In *International Conference on Machine Learning*, 2642–2651. PMLR.

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

Roweis, S. T.; and Saul, L. K. 2000. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500): 2323–2326.

Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; and Chen, X. 2016. Improved techniques for training GANs. *Advances in Neural Information Processing Systems*, 29.

Saxena, D.; and Cao, J. 2021. Generative adversarial networks (GANs) challenges, solutions, and future directions. *ACM Computing Surveys (CSUR)*, 54(3): 1–42.

Schmidhuber, J. 1990. Making the world differentiable: On using fully recurrent self-supervised neural networks for dynamic reinforcement learning and planning in non-stationary environments. *Institut für Informatik, Technische Universität München. Technical Report FKI-126*, 90.

Schmidhuber, J. 1991. A possibility for implementing curiosity and boredom in model-building neural controllers. In *Proc. of the International Conference on Simulation of Adaptive Behavior: From Animals to Animats*, 222–227.

Schmidhuber, J. 2020. Generative adversarial networks are special cases of artificial curiosity (1990) and also closely related to predictability minimization (1991). *Neural Networks*, 127: 58–66.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the Inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2818–2826.

Thanh-Tung, H.; Tran, T.; and Venkatesh, S. 2019. Improving generalization and stability of generative adversarial networks. *International Conference on Learning Representations*.

Xiangli, Y.; Deng, Y.; Dai, B.; Loy, C. C.; and Lin, D. 2020. Real or not real, that is the question. *International Conference on Learning Representations*.

Xie, J.; Lu, Y.; Gao, R.; Zhu, S.-C.; and Wu, Y. N. 2018a. Cooperative training of descriptor and generator networks. *IEEE transactions on pattern analysis and machine intelligence*, 42(1): 27–45.

Xie, J.; Lu, Y.; Zhu, S.-C.; and Wu, Y. 2016. A theory of generative convnet. In *International Conference on Machine Learning*, 2635–2644. PMLR.

Xie, J.; Zheng, Z.; Gao, R.; Wang, W.; Zhu, S.-C.; and Wu, Y. N. 2018b. Learning descriptor networks for 3d shape synthesis and analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8629–8638.

Xie, J.; Zhu, S.-C.; and Nian Wu, Y. 2017. Synthesizing dynamic patterns by spatial-temporal generative convnet. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7093–7101.

Xie, J.; Zhu, S.-C.; and Wu, Y. N. 2019. Learning energy-based spatial-temporal generative convnets for dynamic patterns. *IEEE transactions on pattern analysis and machine intelligence*, 43(2): 516–531.

Yu, J.; Lin, Z.; Yang, J.; Shen, X.; Lu, X.; and Huang, T. S. 2018. Generative image inpainting with contextual attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5505–5514.

Zhao, J.; Mathieu, M.; and LeCun, Y. 2017. Energy-based generative adversarial network. *International Conference on Learning Representations*.