InfoMax-GAN: Improved Adversarial Image Generation via Information Maximization and Contrastive Learning

Kwot Sin Lee[1,2] Ngoc-Trung Tran[3] Ngai-Man Cheung[3]
[1]University of Cambridge [2]Snap Inc.
[3]Singapore University of Technology and Design

Abstract

While Generative Adversarial Networks (GANs) are fundamental to many generative modelling applications, they suffer from numerous issues. In this work, we propose a principled framework to simultaneously mitigate two fundamental issues in GANs: catastrophic forgetting of the discriminator and mode collapse of the generator. We achieve this by employing for GANs a contrastive learning and mutual information maximization approach, and perform extensive analyses to understand sources of improvements. Our approach significantly stabilizes GAN training and improves GAN performance for image synthesis across five datasets under the same training and evaluation conditions against state-of-the-art works. In particular, compared to the state-of-the-art SSGAN, our approach does not suffer from poorer performance on image domains such as faces, and instead improves performance significantly. Our approach is simple to implement and practical: it involves only one auxiliary objective, has low computational cost, and performs robustly across a wide range of training settings and datasets without any hyperparameter tuning. For reproducibility, our code is available in the open-source GAN library, Mimicry [34].

1. Introduction

The field of generative modelling has witnessed incredible successes since the advent of Generative Adversarial Networks (GANs) [16], a form of generative model known for its sampling efficiency in generating high-fidelity data [45]. In its original formulation, a GAN is composed of two models - a generator and a discriminator - which together play an adversarial minimax game that enables the generator to model the true data distribution of some empirical data. This adversarial game is encapsulated by the following equation:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_r(x)}[\log D(x)] + \mathbb{E}_{z\sim p_(z)}[\log(1 - D(G(z)))] \tag{1}$$

where V is the value function, $p(z)$ is a prior noise distribution, $p_r(x)$ is the real data distribution, and $G(z)$ is the generated data from sampling some random noise z.

In this formulation, training the discriminator and generator with their respective minimax loss functions aims to minimize the Jensen-Shannon (JS) divergence between the real and generated data distributions [16] p_r and p_g respectively. However, GAN training is notoriously difficult. Firstly, such theoretical guarantees only come under the assumption of the discriminator being trained to optimality [20], which may lead to saturating gradients in practice. Even so, there is no guarantee for convergence in this minimax game as both generator and discriminator are simultaneously and independently finding a Nash equilibrium in a high-dimensional space. Finally, GANs face the perennial problem of mode collapse, where p_g collapses to only cover a few modes of p_r, resulting in generated samples of limited diversity. Consequently, recent years have seen concerted efforts [15, 23, 46, 54, 59, 60, 62] to mitigate these issues.

A primary cause of GAN training instability is the non-stationary nature of the training environment: as the generator learns, the modeled distribution p_g the discriminator faces is ever changing. As our GAN models are neural networks, the discriminator is susceptible to *catastrophic forgetting* [10, 25, 28, 40], a situation where the network learns ad-hoc representations and forgets about prior tasks to focus on the current one as the weights of the network updates, which contributes to training instability. The state-of-the-art Self-supervised GAN (SSGAN) [10] is the first to demonstrate that a representation learning approach could mitigate discriminator catastrophic forgetting, thus improving training stability. However, the approach still does not explicitly mitigate *mode collapse*, and has a failure mode in image domains such as faces [10]. Furthermore, [61] shows that while SSGAN's approach is helpful for discriminator forgetting, it in fact promotes mode collapse for the generator.

To overcome these problems, we present an approach to simultaneously mitigate both catastrophic forgetting and mode collapse. On the discriminator side, we apply mutual information maximization to improve long-term representation learning, thereby reducing catastrophic forgetting in the non-stationary training environment. On the generator side, we employ contrastive learning to incentivize the generator to produce diverse images that give easily distinguishable positive/negative samples, hence reducing mode collapse. Through mitigating both issues, we show a wide range of practical improvements on natural image synthesis using GANs. We summarize our contributions below:

- We present a GAN framework to improve natural image synthesis through simultaneously mitigating two key GAN issues using just one objective: catastrophic forgetting of the discriminator (via information maximization) and mode collapse of the generator (via contrastive learning). Our approach mitigates issues in both discriminator and generator, rather than either alone.
- With this multi-faceted approach, we significantly improve GAN image synthesis across *five* different datasets against state-of-the-art works under the *same training and evaluation conditions*.
- Our framework is lightweight and practical: it introduces just one auxiliary objective, has a low computational cost, and is robust against a wide range of training settings *without any tuning* required.
- Our work is the first to demonstrate the effectiveness of contrastive learning for significantly improving GAN performance, and also the first to apply the InfoMax principle in a GAN setting, which we hope would open a new research direction in these areas.

2. Background

Mutual information and representation learning Mutual information has deep connections with representation learning [5], where we aim to learn an encoder function E that ideally captures the most important features of the input data X, often at a lower dimensional latent space. This concept is encapsulated by the InfoMax objective [35]:

$$\max_{E \in \mathcal{E}} \mathcal{I}(X; E(X)) \tag{2}$$

where $\mathcal{E}$ is some function class, and the objective is to find some E that maximizes the mutual information between the input data and its encoded representations $E(X)$. To maximize on the InfoMax objective, one could alternatively maximize $\mathcal{I}(C_\psi(X); E_\psi(X))$, where C_ψ and E_ψ are encoders part of the same architecture parameterised by ψ. It is shown in [63] maximizing $\mathcal{I}(C_\psi(X); E_\psi(X))$ is maximizing on a lower bound of the InfoMax objective:

$$\mathcal{I}(C_\psi(X); E_\psi(X)) \leq \mathcal{I}(X; (C_\psi(X), E_\psi(X))) \tag{3}$$

In practice, maximizing $\mathcal{I}(C_\psi(X); E_\psi(X))$ has several advantages: (a) Using different feature encodings allow us to capture different views and modalities of the data for flexibility of modelling [3,57]; (b) The encoded data lies in a much lower dimensional latent space than that of the original data, thus reducing computational constraints [51,63].

Contrastive learning Recently, state-of-the-art works in unsupervised representation learning [3,22,24,29,37,47,57] lies in taking a contrastive approach to maximizing the mutual information between encoded local and global features. Yet, since directly maximizing mutual information is often intractable in practice [49], these works often maximize on the InfoNCE [47] lower bound instead, which involves a contrastive loss minimized through having a critic find positive samples in contrast to a set of negative samples. Such positive/negative samples are arbitrarily created by pairing features [24], augmentation [9], or a combination of both [3]. Our work similarly maximizes on this InfoNCE bound, and most closely follows the Deep InfoMax [24] approach of obtaining local and global features for the maximization.

3. InfoMax-GAN

3.1. Approach

Figure 1 illustrates the InfoMax-GAN framework. Firstly, to maximize on the lower bound of the InfoMax objective, $\mathcal{I}(C_\psi(X); E_\psi(X))$, we set E_ψ to represent layers of the discriminator leading to the global features, and C_ψ as layers leading to the local features. Here, $C_\psi = C_{\psi,1} \circ ... \circ C_{\psi,n}$ is a series of n intermediate discriminator layers leading to the last local feature map $C_\psi(x)$ and f_ψ is the subsequent layer transforming $C_\psi(x)$ to a global feature vector $E_\psi(x)$, which is ultimately used for computing the GAN objective $\mathcal{L}_{\text{gan}}$. We set the local and global feature as the penultimate and final feature outputs of the discriminator encoder respectively, and we study its ablation impact in Appendix A.5.

Next, the local/global features $C_\psi(x)$ and $E_\psi(x)$ extracted from the discriminator are passed to the critic networks ϕ_θ and ϕ_ω to be projected to a higher dimension Reproducing Kernel Hilbert Space (RKHS) [2], which exploits the value of linear evaluation in capturing similarities between the global and local features. These projected features then undergo a Contrastive Pairing phase to create positive/negative samples, where given some image x, a positive sample is created by pairing the (projected) global feature vector $\phi_\omega(E_\psi(x))$ with a (projected) local spatial vector $\phi_\theta(C_\psi^{(i)}(x))$ from the image's own (projected) local feature map $\phi_\theta(C_\psi(x))$, where $i \in \mathcal{A} = \{0, 1, ..., M^2 - 1\}$ is an index to the $M \times M$ local feature map. Doing so, we represent a positive sample as the pair $(\phi_\theta(C_\psi^{(i)}(x)), \phi_\omega(E_\psi(x)))$ for some i. For each of such positive sample, negative samples are obtained by sampling local spatial vectors from the projected local feature map of another image x' in the same mini-batch, and are represented as the pairs $(\phi_\theta(C_\psi^{(i)}(x')), \phi_\omega(E_\psi(x)))$. Intuitively, this step constrains the discriminator to produce global features of some image that maximizes mutual information with the local features of the same image, rather than those from other images.

Taking this further, consider for each positive sample, the

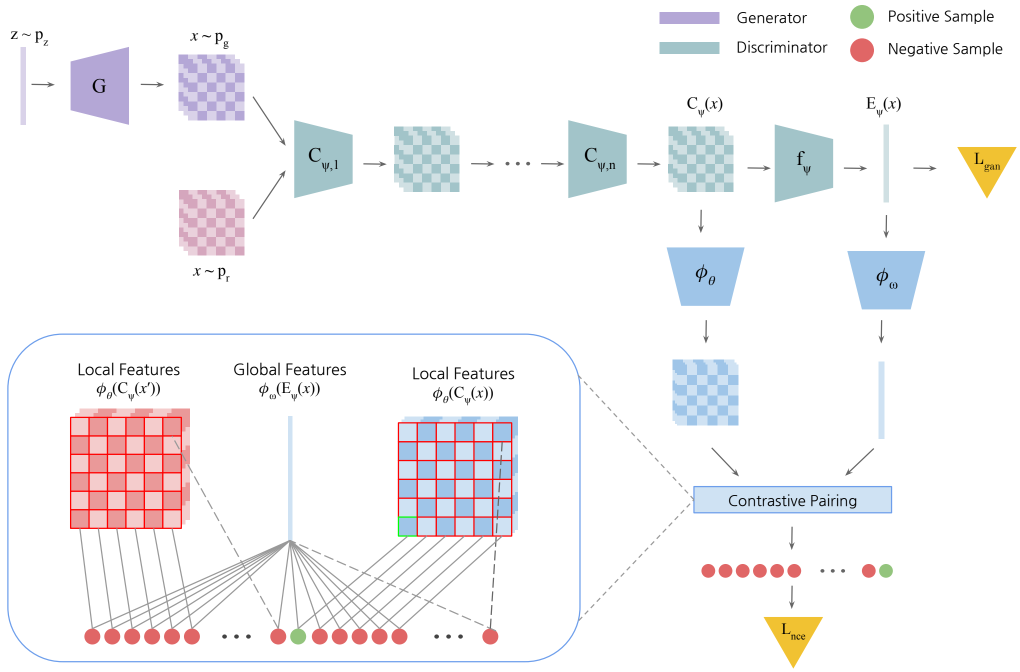


Figure 1: Illustration of the InfoMax-GAN framework. An image x is sampled from the real data distribution p_r or fake data distribution p_g as modeled by the generator G. Image x passes through a discriminator encoder $E_\psi = f_\psi \circ C_\psi$, where $C_\psi = C_{\psi,1} \circ ... \circ C_{\psi,n}$ is a series of n intermediate discriminator layers leading to the last local feature map $C_\psi(x)$ and f_ψ transforms $C_\psi(x)$ to a global feature vector $E_\psi(x)$, which is subsequently used to compute the GAN objective $\mathcal{L}_{\text{gan}}$. The local and global features $C_\psi(x)$ and $E_\psi(x)$ are then projected to a higher dimension by the spectral normalized critic networks ϕ_θ and ϕ_ω respectively. Finally, the resulting features undergo a Contrastive Pairing phase involving local features from another image x', to produce positive and negative samples for computing the contrastive loss $\mathcal{L}_{\text{nce}}$.

pairs $(\phi_\theta(C_\psi^{(j)}(x)), \phi_\omega(E_\psi(x)))$, $j \in \mathcal{A}, j \neq i$ as negative samples. That is, using spatial vectors from the *same* local feature map to create negative samples. Doing so, we regularize the learnt representations to avoid trivial solutions to the mutual information maximization objective, since the global features are constrained to have consistently high mutual information with *all* spatial vectors of its own local feature map, rather than from only some. This effectively aggregates all local information of the image to represent it.

Thus, for N images in a mini-batch, we produce positive/negative samples to perform an NM^2 way classification for each positive sample. Through this approach, it is shown in [47] one maximizes the InfoNCE lower bound of the mutual information $\mathcal{I}(C_\psi(X); E_\psi(X))$. Formally, for a set of N random images $X = \{x_1, ..., x_N\}$ and set $\mathcal{A} = \{0, 1, ..., M^2 - 1\}$ representing indices of a $M \times M$ spatial sized local feature map, the contrastive loss is:

$$\begin{aligned}\mathcal{L}_{nce}(X) &= -\mathbb{E}_{x \in X}\mathbb{E}_{i \in \mathcal{A}}\left[\log p(C_\psi^{(i)}(x), E_\psi(x) \mid X)\right] \\ &= -\mathbb{E}_{x \in X}\mathbb{E}_{i \in \mathcal{A}}\left[\Delta\right]\end{aligned}$$

$$\Delta = \log \frac{\exp(g_{\theta,\omega}(C_\psi^{(i)}(x), E_\psi(x)))}{\sum_{(x',i) \in X \times \mathcal{A}} \exp(g_{\theta,\omega}(C_\psi^{(i)}(x'), E_\psi(x)))} \quad (4)$$

where $g_{\theta,\omega} : \mathbb{R}^{1\times1\times K} \times \mathbb{R}^{1\times1\times K} \to \mathbb{R}$ is a critic mapping the local/global features with K dimensions to a scalar score. Formally, we define $g_{\theta,\omega}$ to be:

$$g_{\theta,\omega}(C_\psi^{(i)}(x), E_\psi(x)) = \phi_\theta(C_\psi^{(i)}(x))^T \phi_\omega(E_\psi(x)) \quad (5)$$

where $\phi_\theta : \mathbb{R}^{M\times M\times K} \to \mathbb{R}^{M\times M\times R}$ and $\phi_\omega : \mathbb{R}^{1\times1\times K} \to \mathbb{R}^{1\times1\times R}$ are the critic networks parameterized by θ and ω respectively, projecting the local and global features to the higher RKHS. In practice, ϕ_θ and ϕ_ω are defined as shallow networks with only 1 hidden layer following [24], but with spectral normalized weights as well. These shallow networks serve to only project the feature dimensions of the input features, and preserve their original spatial sizes.

To stabilize training, we constrain the discriminator to learn from only the contrastive loss of real image features, and similarly for the generator, from only the contrastive loss of fake image features. We formulate the losses for discriminator and generator $\mathcal{L}_D$ and $\mathcal{L}_G$ as such:

$$\mathcal{L}_G = \mathcal{L}_{\text{gan}}(\hat{D}, G) + \alpha\mathcal{L}_{\text{nce}}(X_g) \quad (6)$$

$$\mathcal{L}_D = \mathcal{L}_{\text{gan}}(D, \hat{G}) + \beta\mathcal{L}_{\text{nce}}(X_r) \quad (7)$$

where α and β are hyperparameters; $\hat{D}$ and $\hat{G}$ represent a fixed discriminator and generator respectively; X_r and X_g

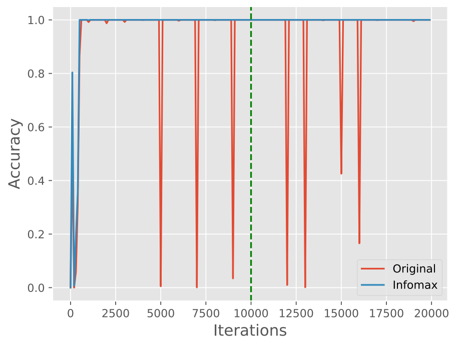


Figure 2: Accuracy of a classifier when trained on the one-vs-all CIFAR-10 classification task. Regularized with the InfoMax objective by minimizing (4), the classifier successfully predicts classes trained from previous iterations even when the underlying class distribution changes.

represent sets of real and generated images respectively; and $\mathcal{L}_{\text{gan}}$ is the hinge loss for GANs [42]:

$$\mathcal{L}_{\text{gan}}(D, \hat{G}) = \mathbb{E}_{x \sim p_r}[\min(0, 1 - D(x))] + \mathbb{E}_{z \sim p_z}[\min(0, 1 + D(\hat{G}(z)))] \tag{8}$$

$$\mathcal{L}_{\text{gan}}(\hat{D}, G) = -\mathbb{E}_{z \sim p_z}[\hat{D}(G(z))] \tag{9}$$

In practice, we set $\alpha = \beta = 0.2$ for all experiments for simplicity, with ablation studies to show our approach is robust across a wide range of α and β values.

3.2. Mitigating Catastrophic Forgetting

Our approach mitigates a key issue in GANs: *catastrophic forgetting* of the discriminator, a situation where due to the non-stationary GAN training environment, the discriminator learns only ad-hoc representations and forget about prior tasks it was trained on. For instance, while the discriminator may learn to penalize flaws in global structures early in GAN training [10], it may later forget these relevant representations in order to learn those for finding detailed flaws in local structures, which overall contributes to training instability.

Inspired by [10], we examine the ability of our approach in mitigating catastrophic forgetting: we train a discriminator classifier on the one-vs-all CIFAR-10 classification task where the underlying class distribution changes every 1K iterations, and the cycle repeats every 10K iterations. As seen in Figure 2, without the InfoMax objective, the classifier overfits to a certain class distribution and produces very low accuracy when the class distribution is changed. When training is regularized with the InfoMax objective, the classifier successfully remembers all prior classes it was trained on. Thus, the InfoMax objective helps the discriminator to reduce catastrophic forgetting and adapt to the non-stationary nature of the generated image distribution, which ultimately stabilizes GAN training.

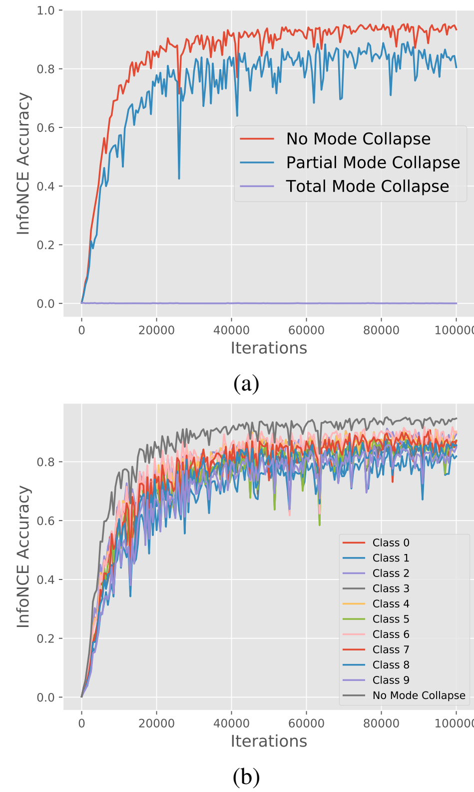


Figure 3: Contrastive task accuracy when simulating generators exhibiting a range of mode collapse behaviours using CIFAR-10 data. (a) We show that the less mode collapsed a generator is, the better the accuracy for contrastive task. (b) The contrastive task accuracy is consistently lower when the generator has partially mode collapsed to any individual class, compared to when there is no mode collapse.

3.3. Mitigating Mode Collapse

Our approach also mitigates a persistent problem of the generator: *mode collapse*. For a fully mode collapsed generator, we have $x = x' \; \forall x, x' \sim X_g$, where X_g is a set of randomly generated images, such that $C_\psi(x) = C_\psi(x')$. This means the term $p(C_\psi^{(i)}(x), E_\psi(x) \mid X_g)$ approaches 0 in the limit, rather than the optimal value 1, as the critics cannot distinguish apart the multiple identical feature pairs from individual images.

To validate this, we show there is a direct correlation between the diversity of generated images and the contrastive learning task accuracy $p(C_\psi^{(i)}(x), E_\psi(x) \mid X)$. We train the discriminator to solve the contrastive task using CIFAR-10 training data, and simulate 3 different kinds of generators using CIFAR-10 test data: (a) a perfect generator with no mode collapse that generates *all classes of images*; (b) a partially mode collapsed generator that only generates *one class of images* and (c) a totally mode collapsed generator that only generates *one image*.

From Figure 3a, we observe a perfect generator with no mode collapse best solves the contrastive task, and a partially mode collapsed generator has a consistently poorer accuracy in the contrastive task than the perfect generator. This concurs with our expectation: images from only one class exhibit a much lower diversity than images from all classes, and so distinguishing positive samples amongst similar and harder negative samples makes the contrastive task much harder. Furthermore, for a totally mode collapsed generator which only generates one image, the accuracy is *near zero*, which confirms our initial hypothesis. For any N images, there are NM^2 samples to classify in the contrastive task, with $NM^2 - 1$ negative samples for each positive sample. However, if all N images are identical due to total mode collapse, then there exists $N - 1$ negative samples identical to each positive sample, which makes the contrastive task nearly impossible to solve. Thus, to solve the contrastive task well, the generator is highly encouraged to generate images with greater diversity, which reduces mode collapse.

Furthermore, in Figure 3b, the performance of any class demonstrating partial mode collapse is *consistently worse* than the case of no mode collapse, where all classes of images are used. Thus, the generator is incentivized to not collapse to producing just any one class that fools the discriminator easily, since producing *all classes* of images naturally leads to the best performance in the contrastive task.

4. Experiments

4.1. Experimental Settings

GAN architectures We compare our model with the baseline Spectral Normalization GAN (SNGAN) [42] and the state-of-the-art Self-supervised GAN (SSGAN) [10]. For clarity, we highlight InfoMax-GAN is equivalent to SNGAN with our proposed objective, and SSGAN is equivalent to SNGAN with the rotation task objective. We show InfoMax-GAN *alone* performs highly competitively, with significant improvements over SSGAN. We detail the exact architectures used for all models and datasets in Appendix C.

Datasets We experiment on five different datasets at multiple resolutions: ImageNet (128×128) [13], CelebA (128×128) [36], CIFAR-10 (32×32) [30], STL-10 (48×48) [12], and CIFAR-100 (32×32) [30]. The details for these datasets can be found in Appendix A.1.

Training We train all models using the same Residual Network [21] backbone, under the *exact* same settings for each dataset, and using the same code base, for fairness in comparisons. For details, refer to Appendix A.2. For all models and datasets, we set $\alpha = \beta = 0.2$, to balance the contrastive loss to be on the same scale as the GAN loss initially. This scaling principle is similar to what is applied in [11], and we later show in our ablation study our framework is highly robust to changes in these hyperparameters.

Evaluation To assess the generated images quality, we employ three different metrics: Fréchet Inception Distance (FID) [23], Kernel Inception Distance (KID) [7], and Inception Score (IS) [54]. In general, FID and KID measure the diversity of generated images, and IS measures the quality of generated images. Here, we emphasize we use the *exact* same number of real and fake samples for evaluation, so that we can compare the scores fairly. This is crucial, especially since metrics like FID can produce highly biased estimates [7], where using a larger sample size leads to a significantly lower score. Finally, for all our scores, we compute them using 3 different random seeds to report their mean and standard deviation. A detailed explanation of all three metrics and the sample sizes used can be found in Appendix A.3

4.2. Results

Improved image synthesis As seen in Table 1, InfoMax-GAN improves FID consistently and significantly across many datasets over SNGAN and SSGAN. On the challenging high resolution ImageNet dataset, InfoMax-GAN improves by **6.8 points** over SNGAN, and **3.6 points** over SSGAN. On the high resolution CelebA, while SSGAN could not improve over the baseline SNGAN, as similarly noted in [10], InfoMax-GAN improves by **3.4 points** over SNGAN, and **5.8 points** over SSGAN. This suggests our approach is versatile and can generalise across multiple data domains.

On STL-10, InfoMax-GAN achieves an improvement of **3.0 points** over SNGAN and **1.5 points** over SSGAN. Interestingly, while InfoMax-GAN performs similarly as SSGAN on CIFAR-10 with around **0.5 points** difference, it improves FID by **3.4 points** on CIFAR-100 when the number of classes increase. We conjecture this is due to the tendency for SSGAN to generate easily rotated images [61], which sacrifices diversity and reduces FID when there are more classes. This observation also supports InfoMax-GAN's larger improvements on ImageNet, which has 1000 classes.

Similarly, for alternative metrics like KID and IS, InfoMax-GAN achieves a highly competitive performance and improves over the state-of-the-art works. On IS, InfoMax-GAN improves from **0.2 to 0.4 points** over SSGAN for all datasets except CIFAR-10, where the margin is less than **0.1 points** and within the standard deviation, indicating a similar performance. Similar to its FID performance on CelebA, SSGAN also performs worse in terms of IS compared to the baseline SNGAN, suggesting its failure mode on faces is not just due to a limited diversity, but also due to poorer quality. In contrast, InfoMax-GAN improves on IS over SNGAN and SSGAN significantly. Finally, on KID, we confirm our result on FID: where FID is better, KID is also

Metric	Dataset	Resolution	Models		
			SNGAN	SSGAN	InfoMax-GAN
FID	ImageNet	128×128	65.74 ± 0.31	62.48 ± 0.31	$\mathbf{58.91 \pm 0.14}$
	CelebA	128×128	14.04 ± 0.02	16.39 ± 0.09	$\mathbf{10.63 \pm 0.04}$
	STL-10	48×48	40.48 ± 0.07	38.97 ± 0.23	$\mathbf{37.49 \pm 0.05}$
	CIFAR-100	32×32	24.76 ± 0.16	24.64 ± 0.16	$\mathbf{21.22 \pm 0.26}$
	CIFAR-10	32×32	18.63 ± 0.22	$\mathbf{16.59 \pm 0.13}$	17.14 ± 0.20
KID	ImageNet	128×128	0.0663 ± 0.0004	0.0616 ± 0.0004	$\mathbf{0.0579 \pm 0.0004}$
	CelebA	128×128	0.0076 ± 0.0001	0.0101 ± 0.0001	$\mathbf{0.0063 \pm 0.0001}$
	STL-10	48×48	0.0369 ± 0.0002	0.0332 ± 0.0004	$\mathbf{0.0326 \pm 0.0002}$
	CIFAR-100	32×32	0.0156 ± 0.0003	0.0161 ± 0.0002	$\mathbf{0.0135 \pm 0.0004}$
	CIFAR-10	32×32	0.0125 ± 0.0001	$\mathbf{0.0101 \pm 0.0002}$	0.0112 ± 0.0001
IS	ImageNet	128×128	13.05 ± 0.05	13.30 ± 0.03	$\mathbf{13.68 \pm 0.06}$
	CelebA	128×128	2.72 ± 0.01	2.63 ± 0.01	$\mathbf{2.84 \pm 0.01}$
	STL-10	48×48	8.04 ± 0.07	8.25 ± 0.06	$\mathbf{8.54 \pm 0.12}$
	CIFAR-100	32×32	7.57 ± 0.11	7.56 ± 0.07	$\mathbf{7.86 \pm 0.10}$
	CIFAR-10	32×32	7.97 ± 0.06	$\mathbf{8.17 \pm 0.06}$	8.08 ± 0.08

Table 1: Mean FID, KID and IS scores of all models across different datasets, computed across 3 different seeds. FID and KID: lower is better. IS: higher is better.

Dataset	Resolution	Models	
		SSGAN	SSGAN + IM
ImageNet	128×128	62.48 ± 0.31	$\mathbf{56.45 \pm 0.29}$
CelebA	128×128	16.39 ± 0.09	$\mathbf{11.93 \pm 0.14}$
STL-10	48×48	38.97 ± 0.23	$\mathbf{37.73 \pm 0.06}$
CIFAR-100	32×32	24.64 ± 0.16	$\mathbf{21.40 \pm 0.20}$
CIFAR-10	32×32	16.59 ± 0.13	$\mathbf{15.42 \pm 0.08}$

Table 2: Mean FID scores (lower is better) of SSGAN before and after applying our method: "+ IM" refers to adding our proposed InfoMax-GAN objective.

better. This further substantiates our FID results and how InfoMax-GAN generates more diverse images across these datasets, with no obvious failure modes unlike in SSGAN.

Orthogonal improvements In Table 2, we show our improvements are orthogonal to those in SSGAN: when adding our objective into SSGAN, FID improves across all datasets significantly, achieving even larger improvements of approximately **2.5 points** for the challenging ImageNet dataset. Thus, our method is flexible and can be easily integrated into existing state-of-the-art works like SSGAN.

Improved training stability Similar to [10], we test training stability through evaluating the sensitivity of model performance when hyperparameters are varied across a range of popular settings for training GANs, such as the Adam parameters (β_1, β_2) and number of discriminator steps per generator step, n_{dis}, all chosen from well-tested settings in seminal GAN works [10, 20, 42, 52, 64]. As seen in Table 3, in comparison to SNGAN at the *same architectural capacity*, InfoMax-GAN consistently improves FID for different datasets even in instances where GAN training does not converge (e.g. when $n_{\text{dis}} = 1$). The FID score variability for InfoMax-GAN is much lower than SNGAN, showing its robustness to changes in training hyperparameters. Finally, while different sets of (β_1, β_2) work better for each dataset, our method stabilizes training and obtain significant improvements in all these settings, *without any hyperparameter tuning*. This can be useful in practice when training new GANs or on novel datasets, where training can be highly unstable when other hyperparameters are not well-tuned.

In Figure 4, we show our method stabilizes GAN training by allowing GAN training to converge faster and consistently improve performance throughout training. We attribute this to an additional constraint where the global features are constrained to have high mutual information with all their local features [24], thereby constraining the space of generated data distribution and causing p_g to change less radically and ultimately stabilizing the GAN training environment. This is a practical benefit when training GANs given a fixed computational budget, since significant improvements can be gained early during training.

Low computational cost In practice, our method takes only a fraction of the training time. Similar to [42], we profile the training time for 100 generator update steps. In Figure 5, we see our approach takes minimal time at less than 0.1% of training time per update, across all n_{dis} for both CIFAR-10 and STL-10. This is since in practice, only 2 shallow (1 hidden layer) MLP networks are needed to

β_1	β_2	n_{dis}	CIFAR-10		STL-10	
			SNGAN	InfoMax-GAN	SNGAN	InfoMax-GAN
0.0	0.9	1	164.74 ± 0.42	$\mathbf{24.42 \pm 0.18}$	267.10 ± 0.20	$\mathbf{54.29 \pm 0.13}$
0.0	0.9	2	20.87 ± 0.19	$\mathbf{18.08 \pm 0.27}$	46.65 ± 0.18	$\mathbf{38.96 \pm 0.31}$
0.0	0.9	5	18.63 ± 0.22	$\mathbf{17.14 \pm 0.20}$	40.48 ± 0.07	$\mathbf{37.49 \pm 0.05}$
0.5	0.999	1	73.07 ± 0.20	$\mathbf{20.58 \pm 0.10}$	134.51 ± 0.37	$\mathbf{62.28 \pm 0.07}$
0.5	0.999	2	18.74 ± 0.24	$\mathbf{17.19 \pm 0.32}$	40.67 ± 0.29	$\mathbf{40.54 \pm 0.20}$
0.5	0.999	5	21.10 ± 0.89	$\mathbf{18.39 \pm 0.04}$	84.20 ± 0.67	$\mathbf{75.72 \pm 0.19}$

Table 3: Mean FID scores (lower is better) across a range of hyperparameter settings. (β_1, β_2) represents the hyperparameters of the Adam optimizer, and n_{dis} represents the number of discriminator steps per generator step. Our method performs robustly in a wide range of training settings *without any tuning*.

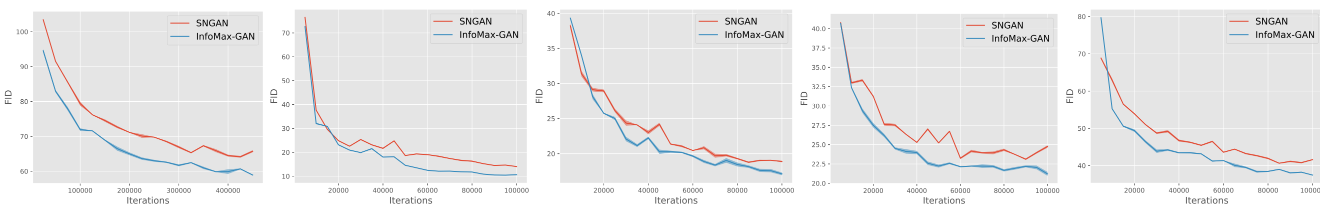
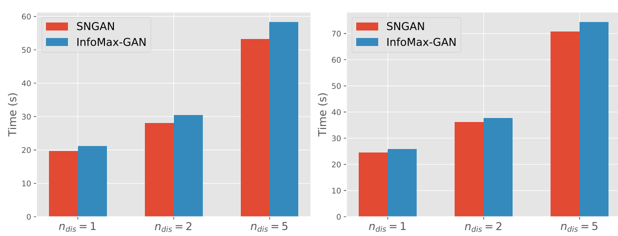


Figure 4: Our approach stabilizes GAN training significantly, converges faster and consistently improves FID for all models across all datasets. Left to right: ImageNet, CelebA, CIFAR-10, CIFAR-100, STL-10.



(a) CIFAR-10 (b) STL-10

Figure 5: Training time for 100 generator update steps across different n_{dis} values for CIFAR-10 and STL-10, using the same hardware. In general, our proposed framework incurs significantly less time than the overall training cost.

compute the contrastive loss. Furthermore, from Table 3, at $n_{\text{dis}} = 2$, InfoMax-GAN has a consistently better FID than SNGAN at $n_{\text{dis}} = 5$ at approximately *half the training time*, since a large n_{dis} is a significant bottleneck in training time. Thus, our approach is practical for training GANs at a fixed computational budget, and has minimal computational overhead.

Improved mode recovery In Appendix A.4, we demonstrate our approach helps to significantly recover more modes in the Stacked MNIST dataset [41].

Qualitative comparisons In Appendix A.6, we show generated images with improved image quality against those

R	Relative Size	FID Score
256	2	$\mathbf{17.07 \pm 0.25}$
512	4	17.21 ± 0.15
1024	8	17.14 ± 0.20
2048	16	17.80 ± 0.05
4096	32	17.38 ± 0.11

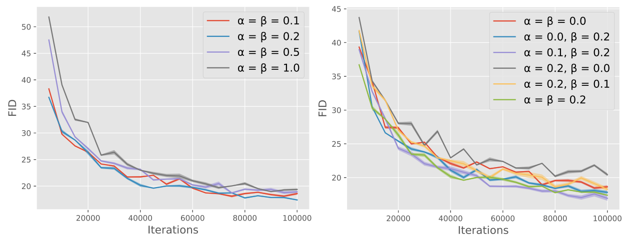
Table 4: Mean FID scores (lower is better) for InfoMax-GAN on CIFAR-10 when the RKHS dimension R is varied. Relative size here refers to how much larger R is relative to the discriminator feature map depth of 128, in terms of multiplicative factor.

from other models for all datasets.

4.3. Ablation Studies

RKHS dimensions As seen in Table 4, our proposed framework is robust to the choice of R, with the FID remaining consistent in their range of values. We attribute this to how the InfoMax critics are simple MLP networks with only 1 hidden layer, which is sufficient for achieving good representations in practice [63]. We note for all our experiments in Tables 1, 2, and 3, we used $R = 1024$.

Sensitivity of α and β hyperparameters In Figure 6a, we performed a large sweep of α and β from 0.0 to 1.0, and see $\alpha = \beta = 0.2$ obtains the best performance for our method. From Figure 6b, we see our InfoMax objective for



(a) (b)

Figure 6: (a) CIFAR-10 FID curves for InfoMax-GAN across a large sweep of α and β hyperparameters, showing $\alpha = \beta = 0.2$ performs the best. (b) We perform a small sweep around the chosen hyperparameters $\alpha = \beta = 0.2$.

the discriminator is important for improving GAN performance: as β is decreased, keeping $\alpha = 0.2$, FID deteriorates. Interestingly, when $\alpha = 0$ and $\beta = 0.2$, having the InfoMax objective for the discriminator alone is sufficient in gaining FID improvements. This confirms our intuition of the role of information maximization in mitigating discriminator catastrophic forgetting to stabilize the GAN training environment and improve FID. However, the performance improves when the generator is also trained on the InfoMax objective, at $\alpha \in \{0.1, 0.2\}$ and $\beta = 0.2$, which affirms our prior intuition that the contrastive nature of the objective helps the generator reduce mode collapse and improve FID. We note apart from this ablation study, we used $\alpha = \beta = 0.2$ for all experiments reported in this paper.

Further studies We include three further ablation studies on our design choices in Appendix A.5 to demonstrate the strength of our approach and justify our design choices.

5. Related Work

Mode collapse and catastrophic forgetting Early works in reducing mode collapse include Unrolled GAN [41], which restructures the generator objective with respect to unrolled discriminator optimization updates. These works often focused on assessing the number of modes recovered by a GAN based on synthetic datasets [8,41,55]. Subsequent works include MSGAN [38], which introduces a regularization encouraging conditional GANs to seek out minor modes often missed when training. These works instead focus on direct metrics [7,19,23,32,53,54] for assessing the diversity and quality of generated images. In our work, we utilized both types of metrics for assessment. Previous approaches to mitigate catastrophic forgetting in GANs include using forms of memory [18,26,54], such as checkpoint averaging. [10] demonstrates the mitigation of catastrophic forgetting using a representation learning approach, which we built upon.

Representation learning and GANs To the best of our knowledge, the closest work in *methodology* to ours is the state-of-the-art SSGAN, which demonstrates the use of a representation learning approach of predicting rotations [14] to mitigate GAN forgetting and hence improve GAN performance. In contrast to SSGAN, our work uses a contrastive learning and information maximization task instead, which we demonstrate to simultaneously mitigate *both* GAN forgetting and mode collapse. Furthermore, our work overcomes failure modes demonstrated in SSGAN, such as in datasets involving faces [10]. For fair and accurate comparisons, our work compared with SSGAN using the *exact* same architectural capacity, training and evaluation settings.

Information theory and GANs The most prominent work in utilizing mutual information maximization for GANs is InfoGAN, but we emphasize here that our work has a *different focus*: while InfoGAN focuses on learning disentangled representations, our goal is to improve image synthesis. For clarity, we illustrate the specific differences with InfoGAN in Appendix B. Other approaches employing information-theoretic principles include Variational GAN (VGAN) [50], which uses an information bottleneck [58] to regularize the discriminator representations; with [6,39,44] extending to minimise divergences apart from the original JS divergence. In contrast to these works, our work employs the InfoMax principle to improve discriminator learning and provides a clear connection to how this improves GAN training via the mitigation of catastrophic forgetting.

6. Conclusion and Future Work

In this paper, we presented the InfoMax-GAN framework for improving natural image synthesis through simultaneously alleviating two key issues in GANs: catastrophic forgetting of the discriminator (via information maximization), and mode collapse of the generator (via contrastive learning). Our approach significantly improves on the natural image synthesis task for *five* widely used datasets, and further overcome failure modes in state-of-the-art models like SSGAN. Our approach is simple and practical: it has only one auxiliary objective, performs robustly in a wide range of training settings without any hyperparameter tuning, has a low computational cost, and demonstrated improvements even when integrated to existing state-of-the-art models like SSGAN. As future work, it would be interesting to explore this framework for different tasks, such as in 3D view synthesis, where one could formulate objectives involving mutual information and adjacent views. To the best of our knowledge, our work is the first to investigate using information maximization and contrastive learning to improve GAN image synthesis performance, and we hope our work opens up new possibilities in this direction.

References

[1] GANs with Spectral Normalization and Projection Discriminator. `https://github.com/pfnet-research/sngan_projection`. Accessed: 2019-10-31.

[2] Nachman Aronszajn. Theory of Reproducing Kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

[3] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning Representations by Maximizing Mutual Information Across Views. *arXiv Preprint arXiv:1906.00910*, 2019.

[4] David Barber and Felix V Agakov. The IM Algorithm: A Variational Approach to Information Maximization. In *Advances in Neural Information Processing Systems*, page None, 2003.

[5] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

[6] Himesh Bhatia, William Paul, Fady Alajaji, Bahman Gharesifard, and Philippe Burlina. R\'{e} nyi Generative Adversarial Networks. *arXiv preprint arXiv:2006.02479*, 2020.

[7] Mikołaj Bińkowski, Dougal J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying Mmd Gans. *arXiv Preprint arXiv:1801.01401*, 2018.

[8] Tong Che, Yanran Li, Athul Paul Jacob, Yoshua Bengio, and Wenjie Li. Mode Regularized Generative Adversarial Networks. *arXiv Preprint arXiv:1612.02136*, 2016.

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations. *arXiv Preprint arXiv:2002.05709*, 2020.

[10] Ting Chen, Xiaohua Zhai, Marvin Ritter, Mario Lucic, and Neil Houlsby. Self-Supervised Gans via Auxiliary Rotation Loss. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12154–12163, 2019.

[11] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, pages 2172–2180, 2016.

[12] Adam Coates, Andrew Ng, and Honglak Lee. An Analysis of Single-Layer Networks in Unsupervised Feature Learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 215–223, 2011.

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A Large-Scale Hierarchical Image Database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009.

[14] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised Representation Learning by Predicting Image Rotations. *arXiv Preprint arXiv:1803.07728*, 2018.

[15] Ian Goodfellow. NIPS 2016 Tutorial: Generative Adversarial Networks. *arXiv Preprint arXiv:1701.00160*, 2016.

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

[17] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A Kernel Two-Sample Test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

[18] Paulina Grnarova, Kfir Y Levy, Aurelien Lucchi, Thomas Hofmann, and Andreas Krause. An Online Learning Approach to Generative Adversarial Networks. *arXiv Preprint arXiv:1706.03269*, 2017.

[19] Paulina Grnarova, Kfir Y Levy, Aurelien Lucchi, Nathanael Perraudin, Ian Goodfellow, Thomas Hofmann, and Andreas Krause. A Domain Agnostic Measure for Monitoring and Evaluating GANs. In *Advances in Neural Information Processing Systems*, pages 12069–12079, 2019.

[20] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved Training of Wasserstein Gans. In *Advances in Neural Information Processing Systems*, pages 5767–5777, 2017.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[22] Olivier J Hénaff, Ali Razavi, Carl Doersch, SM Eslami, and Aaron van den Oord. Data-Efficient Image Recognition with Contrastive Predictive Coding. *arXiv Preprint arXiv:1905.09272*, 2019.

[23] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017.

[24] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning Deep Representations by Mutual Information Estimation and Maximization. *arXiv Preprint arXiv:1808.06670*, 2018.

[25] Ronald Kemker, Marc McClure, Angelina Abitino, Tyler L Hayes, and Christopher Kanan. Measuring Catastrophic Forgetting in Neural Networks. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[26] Youngjin Kim, Minjung Kim, and Gunhee Kim. Memorization Precedes Generation: Learning Unsupervised Gans with Memory Networks. *arXiv Preprint arXiv:1803.01500*, 2018.

[27] Diederik P Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv Preprint arXiv:1412.6980*, 2014.

[28] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming Catastrophic Forgetting in Neural Networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

[29] Lingpeng Kong, Cyprien de Masson d'Autume, Wang Ling, Lei Yu, Zihang Dai, and Dani Yogatama. A Mutual Information Maximization Perspective of Language Representation Learning. *arXiv Preprint arXiv:1910.08350*, 2019.

[30] Alex Krizhevsky, Geoffrey Hinton, et al. Learning Multiple Layers of Features from Tiny Images. Technical report, Citeseer, 2009.

[31] Karol Kurach, Mario Lučić, Xiaohua Zhai, Marcin Michalski, and Sylvain Gelly. A Large-scale Study on Regularization and Normalization in GANs. In *International Conference on Machine Learning*, pages 3581–3590, 2019.

[32] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved Precision and Recall Metric for Assessing Generative Models. In *Advances in Neural Information Processing Systems*, pages 3929–3938, 2019.

[33] Yann LeCun. The MNIST Database of Handwritten Digits. *Http://yann. Lecun. Com/Exdb/Mnist/*, 1998.

[34] Kwot Sin Lee and Christopher Town. Mimicry: Towards the Reproducibility of GAN Research. *CVPR AI for Content Creation Workshop*, 2020.

[35] Ralph Linsker. Self-Organization in a Perceptual Network. *Computer*, 21(3):105–117, 1988.

[36] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[37] Sindy Löwe, Peter O'Connor, and Bastiaan S Veeling. Greedy InfoMax for Biologically Plausible Self-Supervised Representation Learning. *arXiv Preprint arXiv:1905.11786*, 2019.

[38] Qi Mao, Hsin-Ying Lee, Hung-Yu Tseng, Siwei Ma, and Ming-Hsuan Yang. Mode Seeking Generative Adversarial Networks for Diverse Image Synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1429–1437, 2019.

[39] Xudong Mao, Qing Li, Haoran Xie, Raymond Y.K. Lau, Zhen Wang, and Stephen Paul Smolley. Least Squares Generative Adversarial Networks. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[40] Michael McCloskey and Neal J Cohen. Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem. In *Psychology of Learning and Motivation*, volume 24, pages 109–165. Elsevier, 1989.

[41] Luke Metz, Ben Poole, David Pfau, and Jascha Sohl-Dickstein. Unrolled Generative Adversarial Networks. *arXiv Preprint arXiv:1611.02163*, 2016.

[42] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral Normalization for Generative Adversarial Networks. *arXiv Preprint arXiv:1802.05957*, 2018.

[43] Takeru Miyato and Masanori Koyama. cGANs with Projection Discriminator. *arXiv Preprint arXiv:1802.05637*, 2018.

[44] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-GAN: Training Generative Neural Samplers using Variational Divergence Minimization. In *Advances in Neural Information Processing Systems*, pages 271–279, 2016.

[45] Augustus Odena. Open Questions About Generative Adversarial Networks. *Distill*, 2019. https://distill.pub/2019/gan-open-problems.

[46] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional Image Synthesis with Auxiliary Classifier Gans. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2642–2651. JMLR. org, 2017.

[47] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation Learning with Contrastive Predictive Coding. *arXiv Preprint arXiv:1807.03748*, 2018.

[48] Sherjil Ozair, Corey Lynch, Yoshua Bengio, Aaron Van den Oord, Sergey Levine, and Pierre Sermanet. Wasserstein Dependency Measure for Representation Learning. In *Advances in Neural Information Processing Systems*, pages 15604–15614, 2019.

[49] Liam Paninski. Estimation of Entropy and Mutual Information. *Neural Computation*, 15(6):1191–1253, 2003.

[50] Xue Bin Peng, Angjoo Kanazawa, Sam Toyer, Pieter Abbeel, and Sergey Levine. Variational Discriminator Bottleneck: Improving Imitation Learning, Inverse Rl, and Gans by Constraining Information Flow. *arXiv Preprint arXiv:1810.00821*, 2018.

[51] Ben Poole, Sherjil Ozair, Aäron van den Oord, Alexander A Alemi, and George Tucker. On Variational Lower Bounds of Mutual Information. In *NeurIPS Workshop on Bayesian Deep Learning*, 2018.

[52] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *arXiv Preprint arXiv:1511.06434*, 2015.

[53] Mehdi SM Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing Generative Models via Precision and Recall. In *Advances in Neural Information Processing Systems*, pages 5228–5237, 2018.

[54] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved Techniques for Training Gans. In *Advances in Neural Information Processing Systems*, pages 2234–2242, 2016.

[55] Akash Srivastava, Lazar Valkov, Chris Russell, Michael U Gutmann, and Charles Sutton. Veegan: Reducing Mode Collapse in Gans Using Implicit Variational Learning. In *Advances in Neural Information Processing Systems*, pages 3308–3318, 2017.

[56] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.

[57] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive Multiview Coding. *arXiv Preprint arXiv:1906.05849*, 2019.

[58] Naftali Tishby, Fernando C Pereira, and William Bialek. The Information Bottleneck Method. *arXiv Preprint Physics/0004057*, 2000.

[59] Ngoc-Trung Tran, Tuan-Anh Bui, and Ngai-Man Cheung. Dist-Gan: An Improved GAN Using Distance Constraints. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[60] Ngoc-Trung Tran, Tuan-Anh Bui, and Ngai-Man Cheung. Improving GAN with Neighbors Embedding and Gradient Matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5191–5198, 2019.

[61] Ngoc-Trung Tran, Viet-Hung Tran, Bao-Ngoc Nguyen, Linxiao Yang, et al. Self-Supervised GAN: Analysis and Improvement with Multi-Class Minimax Game. In *Advances in Neural Information Processing Systems*, pages 13232–13243, 2019.

[62] Ngoc-Trung Tran, Viet-Hung Tran, Ngoc-Bao Nguyen, and Ngai-Man Cheung. An Improved Self-supervised GAN via Adversarial Training. *arXiv preprint arXiv:1905.05469*, 2019.

[63] Michael Tschannen, Josip Djolonga, Paul K Rubenstein, Sylvain Gelly, and Mario Lucic. On Mutual Information Maximization for Representation Learning. *arXiv Preprint arXiv:1907.13625*, 2019.

[64] David Warde-Farley and Yoshua Bengio. Improving Generative Adversarial Networks with Denoising Feature Matching. 2016.