IID-GAN: an IID Sampling Perspective for Regularizing Mode Collapse

Yang Li, Liangliang Shi and Junchi Yan
Department of Computer Science and Engineering, MoE Key Lab of Artificial Intelligence, Shanghai Jiao Tong University
{yanglily, shiliangliang, yanjunchi}@sjtu.edu.cn

Abstract

Despite its success, generative adversarial networks (GANs) still suffer from mode collapse, i.e., the generator can only map latent variables to a partial set of modes in the target distribution. In this paper, we analyze and seek to regularize this issue with an independent and identically distributed (IID) sampling perspective and emphasize that holding the IID property referring to the target distribution for generation can naturally avoid mode collapse. This is based on the basic IID assumption for real data in machine learning. However, though the source samples \( \{z\} \) obey IID, the generations \( \{G(z)\} \) may not necessarily be IID sampling from the target distribution. Based on this observation, considering a necessary condition of IID generation that the inverse samples from target data should also be IID in the source distribution, we propose a new loss to encourage the closeness between inverse samples of real data and the Gaussian source in latent space to regularize the generation to be IID from the target distribution. Experiments on both synthetic and real-world data show the effectiveness of our model.

1 Introduction

For generative models, the training data (in target space) is often assumed to be IID sampled from an unknown implicit distribution. Deep generative models often try to construct a mapping from a known distribution in source space (e.g. Gaussian distribution) to the implicit target distribution. Among popular generative schemes, e.g. Variational Auto-Encoder (VAEs) [Kingma and Welling, 2014], Generative Flow models [Rezende and Mohamed, 2015] and Generative Adversarial Networks (GANs) [Goodfellow et al., 2014], GAN is widely recognized for its excellence in generating images with high resolution. Mode collapse is one of the standing issues in GAN, a phenomenon that the generator tends to get stuck in a subset of modes while excluding other parts of the target distribution [Liu et al., 2020; Yang et al., 2019], leading to poor generation diversity.

Efforts have been made to address mode collapse along two branches: 1) Achieve a better convergence between the generated distribution and the target (real data) distribution [Gulrajani et al., 2017; Metz et al., 2017]. Though the distribution convergence is recognized as the ultimate goal of generation [Goodfellow et al., 2014], the implicitness of the two distributions makes the desire can only be achieved by utilizing the sampling data. Thus the goal of distribution convergence can be somehow vague and imprecise. 2) Penalize the similarity of the generated images [Elfeki et al., 2019; Mao et al., 2019] or apply multiple generator/discriminators [Liu and Tuzel, 2016; Nguyen et al., 2017]. These methods promote the generation diversity in a more direct manner but often lack rigorous theoretical guarantees.

Existing mainstream GAN methods [Goodfellow et al., 2014; Arjovsky et al., 2017; Gulrajani et al., 2017; Miyato et al., 2018] focus on achieving the identical generated distribution as the real one. [Goodfellow et al., 2014] shows that if given enough capacity and training time, the generated distribution converges to real distribution through GAN’s optimization. However, due to the limited capacity of the networks (e.g. catastrophic forgetting) and the implicitness of the two distributions, the convergence only holds in theory. With the short-term guidance in training, most of the methods [Goodfellow et al., 2014; Arjovsky et al., 2017; Gulrajani et al., 2017; Miyato et al., 2018] tend to sample-wise boost the generation quality i.e. increase the probability density of the generated sample obeying the target distribution. Taking GAN as an example, discriminator \( D \) distinguishes whether every generation is sampled from the identical target distribution, and generator \( G \) tends to generate samples with high discriminative confidence of the current \( D \). While in VAE-based or Flow-based approaches [Kingma and Welling, 2014; Rezende and Mohamed, 2015], log-likelihood is applied to increase the density of every generation. These methods rarely study the coupled relation between the generated samples to constitute a distribution, and each step the optimized sample is pulled to the center of the distribution to maintain high probability density, while the characteristics of the overall distribution e.g. variance are not explicitly controlled. This leads them to merely improve the quality of the individual generated samples, trivializing to maintain diversity as real data does. This may stem from the dilemma that a single sample can hardly estimate a distribution through nets, unless the nets possess unrestricted capability and do not exhibit catastrophic forgetting [McCloskey and Cohen, 1989].

In this paper, rather than study the relation between distribu-
tions, we directly optimize the sampling relation between generated sample batches and the target distribution. We first make a basic yet under-exploited (in GAN literature) observation that datasets for generation are assumed to be independent and identically distributed (IID) sampled from an unknown target distribution (i.e., real distribution). We aim to achieve that the generated samples are independently sampled from the identical target distribution. If the latents \( \{z\} \) are IID sampled from the source, it is guaranteed that the generations \( \{G(z)\} \) are IID samples from the generated distribution, but not necessarily IID from the target. Note the independence here is considered from the perspective of statistical test, which needs consistency over samples \( \{x_1, x_2, \cdots, x_n\} \) and the target distribution (or between the latents \( \{z_1, z_2, \cdots, z_n\} \) and source distribution). Sample batches can contain relatively more complete features characterizing a distribution, improving optimization efficiency.

To address mode collapse and get IID generation, we first define mode completeness as the key requirement of the ideal mapping from the source to the target. As the target distribution is implicit and we can only obtain its finite IID samples, we propose a weaker/necessary condition\(^1\) for IID generation based on an inverse mapping: if the real data are IID samples from target distribution, then their inverses in the source space are also IID. To achieve IID property, a common and straightforward idea is to drive the distribution of the overall inverse samples close to a standard Gaussian by certain measures. This idea differs from existing inversion-based methods [Srivastava et al., 2017; Donahue et al., 2017; Yu et al., 2019] which learn the relationship between the individual latent sample \( z \) and real data \( x \) through the discriminator, yet the relationship among \( \{z\} \) is not fully exploited. The highlights of the paper are:

1) We take an IID perspective to mode collapse in GAN. The requirement for an ideal generator is proposed that it should satisfy the so-called mode completeness by Def. 1. Limited by the finiteness of real samples, we resort to a necessary condition of mode completeness in Prop. 1, which requires the IID property of the inverse samples from the real data.

2) Guided by Prop. 1, a regulator is devised to avoid mode collapse. We enforce the inverse sample batch to be distributed close to a standard Gaussian by Wasserstein distance, termed as Gaussian consistency loss (see Sec. 4.3). QQ-plot, Shapiro–Wilks and Kolmogorov-Smirnov statistics are used to test the IID property of the inverse Gaussian samples.

3) We show that IID-GAN outperforms baselines by different metrics on synthetic data w.r.t. the number of covered modes, quality and reverse KL divergence. Our regularization technique also performs competitively on natural images. Unsupervised disentanglement feature learning and conditional generation are also investigated.

2 Related Work

Since its debut [Goodfellow et al., 2014], subsequent works of GANs have emerged to improve training stability and generation quality. However, issues still remain for GANs, and mode collapse is one of the most common challenges. In this section, we introduce methods that are directly or indirectly dedicated to address mode collapse.

Improving training behavior. The promotion of distribution convergence can naturally avoid mode collapse. Unrolled GAN [Metz et al., 2017] presents a surrogate objective to train the generator, along with an unrolled optimization of the discriminator to improve training stability. Wasserstein GANs [Arjovsky et al., 2017] and its variants e.g. [Gulrajani et al., 2017] modify the optimization objective to stabilize training. Although these methods all claim to improve the convergence between the generated and real distribution, the convergence focus more on generation quality that they tend to align every generated sample closer to the real distribution but focus less on the inter-sample relation.

Enforcing to capture diverse modes. Many other methods address mode collapse by penalizing similarity. GDPP [Elfeki et al., 2019] applies the determinantal point process theory, which imposes a penalty on the discriminator to enforce the convergence of the covariance matrix between the features of generated samples and real data. BuresGAN [De Meulmeester et al., 2021] matches the discriminator’s last layer output through Bures distance. AdvLatGAN [Li et al., 2022] mines sample pairs that tend to collapse and regularize them to maintain diversity. All these methods are concerned with the relationship among generated results. However, these methods lack stability guarantees for diverse generation, as they ignore that the generated data shall statistically maintain the same diversity as the real distribution.

Multiple generators and discriminators. Involving multiple generators to achieve wider coverage over the real distribution can also benefit diverse generation. In [Liu and Tuzel, 2016], two coupled generators are trained with parameter sharing to learn the real distribution. The multi-agent system MAD-GAN [Ghosh et al., 2018] involves multiple generators along with one discriminator, and it explicitly encourages each generator to learn a partial set of modes. On the other hand, multiple discriminators are used in [Durugkar et al., 2017] as an ensemble. AMAT [Mangalam and Garg, 2021] introduce a novel training procedure that adaptively spawns additional discriminators to remember previous modes of generation. These methods directly seek to utilize more networks to capture more modes, but omit to reach the essence of mode collapse.

Learn the representations by inverse mapping. BigAN [Donahue et al., 2017], VEEGAN [Srivastava et al., 2017], VAE-based models [Kingma and Welling, 2014] and recently proposed MGGAN [Bang and Shim, 2021], DistGAN [Tran et al., 2018] involve encoding networks (inverse mapping) of the generator to encourage the convergence between the inverse distribution of real data and source distribution. Flow-GAN [Grover et al., 2018] utilizes a flow-based generator that allows for invertibility. However, it requires very special network design that can significantly limit the network’s expressive power and inference speed.

However, similar to the previous discussion about methods improving training behavior, the inverse mapping procedure ignores the IID requirement, i.e., they merely sample-wise boost the generation quality (e.g. the probability of the sample locating in real data distribution) and fail to explicitly control

\(^1\) Note that it will become a necessary and sufficient condition if there are an infinite amount of real data.
We first define mode completeness. So the mode completeness can also be interpreted as SB where space B ensures the probability measures for set-set pairs in the two spaces should be equal. (a) Mode collapse caused by real data’s inverses outside high-density sampling region. 

Figure 1: Mapping from source to target: (a) Left: high-density samples can only be mapped to part of real data i.e. mode collapse; Right: only part of real samples’ inverses lie in high-density sampling region, and the inverses outside the region can hardly be sampled, leading to mode collapse. (b) Assuming the existence of the inverse of mapping G(·), to solve mode collapse and get IID generation, the probability measures of the set G⁻¹(S) and S should be equal for any set S in target space. See Fig. 3 for experimental examples.

We naturally avoid the mass of the probability measure Gf(x) = 1, 2, · · · , n because the values of the probability measures given the corresponding sets based on G are equal (i.e. α({z_i}) = β({T(z_i)})). Then given the equal probability measures, when z_i are IID sampled from α, the generations {G(z_i)} are IID samples from β.

Mode completeness β = G#α can hardly be completely achieved since β is complex and implicit. Considering the basic assumption that the real training data are IID sampled from β, if we assume the existence of G⁻¹ : B → A such that z = G⁻¹(G(z)) and x = G(G⁻¹(x)) for any z ∈ A and x ∈ B, then we can obtain a necessary condition for mode completeness. Such a necessary condition for IID generation is formalized Proposition 1:

**Proposition 1. (IID for Inverse of Target Samples)** If the mapping G satisfies mode completeness from the source probability measure α to the target measure β and its inverse G⁻¹ exists, then given IID samples {x(i)}n i=1 from β, their inverses {G⁻¹(x(i))}n i=1 can be viewed as IID samples from α.

To prove Proposition 1, we first provide a proof for Lemma 1, which represents a more general scenario.

**Lemma 1. Assume the existence of mapping T which satisfies T#α = β and the existence of its inverse T⁻¹.** Let ρ be the joint probability measure of n independent copies of β, denoted as β₁, · · · , βₙ and π is the joint probability measure of n independent copies of α, denoted as α₁, · · · , αₙ. Then we have T#π = ρ, where T is the concatenation of n mapping Ts that satisfies T(x₁, · · · , xₙ) = [T(x₁), · · · , T(xₙ)]. Then we can obtain that π is the joint probability measure with n independent probability measure α₁, · · · , αₙ.

**Proof.** Given n measurable set Sᵢ ⊂ B and S = S₁ × S₂ × · · · × Sₙ, with the independence between {βᵢ}n i=1, we obtain:

\[ ρ(S) = β₁(S₁)β₂(S₂) · · · βₙ(Sₙ) \]  

Since T#π = ρ and T#α = β, we know that ρ(S) = π(T⁻¹(S)) and β(S) = α(T⁻¹(S₁)). Then we have:

\[ π(T⁻¹(S₁)) = α₁(T⁻¹(S₁)) · · · αₙ(T⁻¹(Sₙ)) \]

which indicates that π is the joint probability measure with independent α₁, · · · , αₙ.

The proof of Proposition 1 can be easily obtained from Lemma 1 by setting n measurable sets as Sᵢ = {x₁}, · · · , Sₙ = {xₙ} where x₁, · · · , xₙ are n IID samples from target distribution. According to Eq. 3, we obtain

\[ α(T⁻¹(x₁)) · · · α(T⁻¹(xₙ)) = π(T⁻¹({x₁, · · · , xₙ})) \]

which implies that {T⁻¹(x(i))}n i=1 can be viewed as n independent samples from source distribution.

Since it is commonly assumed that the real data {x(i)}n i=1 are independently sampled from an unknown distribution β, their inverses {G⁻¹(x(i))}n i=1 can be viewed as IID samples from a known distribution α. To satisfy mode completeness,
we regularize the inverse samples of real data to be closer to the IID samples from \( \alpha \). Since we can hardly obtain a strict inverse mapping \( G^{-1} \), we call for an approximation of \( G^{-1} \) as \( F \), which is obtained by penalizing:

\[
\mathbf{z} = F(G(\mathbf{x})) \quad \mathbf{x} = G(F(\mathbf{z})). \tag{5}
\]

### 4 The Proposed Approach

Our loss mainly consists of three parts as shown in Fig. 2.

#### 4.1 Adversarial Learning Term: GAN Loss

The vanilla GAN [Goodfellow et al., 2014] consists of a discriminator \( D : \mathcal{R}^d \to \mathcal{R} \) and a generator \( G : \mathcal{R}^M \to \mathcal{R}^d \), which are typically embodied by deep neural networks. Given the empirical distribution \( p(x) \), \( D \) distinguishes whether the generated sample is from real data, while \( G \) models the mapping from Gaussian sample \( \mathbf{z} \) to a target space sample. The objective \( V(G, D) \) is optimized for the discriminator and generator by solving the min-max:

\[
E_{\mathbf{x} \sim p_r(\mathbf{x})} [\log(1 - D(G(\mathbf{x}))) + E_{\mathbf{z} \sim p(\mathbf{z})} [\log(D(G(\mathbf{z})))]] \quad \tag{6}
\]

The first term denotes the probability expectation of \( \mathbf{x} \) coming from the real data distribution \( p(\mathbf{z}) \) and the second term involves the input distribution \( p(\mathbf{z}) \), which is embodied in this paper as a standard multi-dimensional \((M-D)\) Gaussian distribution \( \mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I}) \). \( \mathbf{I} \in \mathcal{R}^{M \times M} \) denotes the identity matrix.

#### 4.2 Reconstruction Term: Cycle-Consistency Loss

Prop. 1 calls for the existence of the generator’s inverse, i.e. \( G^{-1} \). To achieve the approximation for the holding of \( \mathbf{z} = F(G(\mathbf{z})) \) and \( \mathbf{x} = G(F(\mathbf{x})) \) where \( F \) is also a network-based mapping, a cycle-consistency loss can be an effective solution [Srivastava et al., 2017; Donahue et al., 2017]:

\[
L_{\text{re}}(G, F) = E_{\mathbf{z} \sim p(\mathbf{z})} \| \mathbf{z} - F(G(\mathbf{z})) \|_2 + \lambda E_{\mathbf{x} \sim p(\mathbf{x})} \| \mathbf{x} - G(F(\mathbf{x})) \|_2 \quad \tag{7}
\]

Here \( \lambda = d/M \) is the dimension ratio of \( \mathbf{x} \) to \( \mathbf{z} \). The first term promotes \( F \) to be the inverse of \( G \), which takes the reconstruction loss as the expectation of the cost for auto-encoding noise vectors [Srivastava et al., 2017]. The second term promotes \( F(\mathbf{x}) \in \mathcal{R}^M \), which makes \( \mathbf{z} = F(\mathbf{x}) \) possible to be sampled from Gaussian. Then for each \( \mathbf{x} \), we can obtain the corresponding \( \hat{\mathbf{z}} \in \mathcal{R}^M \) satisfying \( G(\hat{\mathbf{z}}) = \mathbf{x} \), which guarantees the existence of the real samples’ inverses in source space. Note that it does not offer any help to avoid mode imbalance, since the imbalance refers to the overall distribution rather than the individual data points studied here.

#### 4.3 Regularizer: Inverse Distribution Consistency

Recall the necessary condition for IID generation proposed in Prop. 1: assume that the given real data \( \{\mathbf{x}^{(i)}\}_{i=1}^{n} \) are IID sampled from \( p(\mathbf{x}) \), then the inverse samples of the real data will be independent and obey the same distribution \( p(\mathbf{z}) \) (i.e. IID samples from source distribution). Thus, given a batch of real data, \( \{F(\mathbf{x}^{(i)})\}_{i=1}^{n} \) should be closer to independent samples from standard Gaussian \( p(\mathbf{z}) \).

**M-D Gaussian consistency loss.** Suppose the inverse samples \( \{\tilde{\mathbf{z}}^{(i)}\}_{i=1}^{n} \in \mathcal{R}^M \) of real data obey the Gaussian distribution \( \mathcal{N}(\mathbf{z}; \hat{\mu}, \hat{\Sigma}) \) in latent space, then the maximum likelihood estimation can be formulated as:

\[
\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} \tilde{\mathbf{z}}^{(i)} \quad \hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (\tilde{\mathbf{z}}^{(i)} - \hat{\mu})^\top (\tilde{\mathbf{z}}^{(i)} - \hat{\mu}) \quad \tag{8}
\]

where \( \hat{\mu} \in \mathcal{R}^M \), and \( \hat{\Sigma} \in \mathcal{R}^{M \times M} \) is the estimated covariance. Our goal is to make the Gaussian \( q(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \hat{\mu}, \hat{\Sigma}) \) closer to the standard Gaussian \( p(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I}) \), based on which we can view \( \{\tilde{\mathbf{z}}^{(i)}\}_{i=1}^{n} \) as IID samples from standard Gaussian \( p(\mathbf{z}) \). To this end, we introduce a distribution closeness loss \( L_{\text{Gauss}} \). For example, it can be specified as the square of Wasserstein distance of two Gaussians:

\[
L_{\text{Gauss}} = \|\hat{\mu}\|^2 + \text{trace}(\hat{\Sigma} + \mathbf{I} - 2\hat{\Sigma}^{1/2}) \quad \tag{9}
\]

The two Gaussian distributions \( q(\mathbf{z}) \) and \( p(\mathbf{z}) \) can also be evaluated by static divergence, e.g., \( p \)-norm, KL-divergence. More details about loss selections are presented in Appendix A.

**Decoupling M-D Gaussian into M 1-D Gaussians.** For large \( M \) and small batch size, the training may potentially suffer from the curse of dimensionality for estimating \( \hat{\Sigma} \). So we also devise a simplified consistency loss by decoupling the \( M-D \) Gaussian loss to the sum of the \( 1-D \) Gaussian losses of \( M \) ones. This design introduces an assumption on the covariance that the non-diagonal values are all zero. Equivalently, the inverse samples \( \{\tilde{\mathbf{z}}^{(i)}\}_{i=1}^{n} \) follow \( M \) 1-D Gaussian \( \mathcal{N}(\mathbf{z}_i; \mu_i, \sigma_i) \) and then we can get the estimation \( \hat{\mu} \) and \( \hat{\sigma} \).

Our goal is to make the 1-D Gaussian distribution \( q(\mathbf{z}_i) \) closer to the standard 1-D Gaussian distribution in each dimension \( j \). Similar to the \( M-D \) case, we can design the Gaussian loss with Wasserstein distance between two Gaussian distributions, and summing over \( M \) dimensions:

\[
L_{\text{Gauss}} = \sum_{m=1}^{M} (\hat{\mu}_m^2 + (\hat{\sigma}_m - 1)^2) \quad \tag{10}
\]

Throughout the rest of this paper, we directly call it the Gaussian loss, omitting the term consistency.

#### 4.4 IID-GAN: an Overview

By inputting samples from \( p(\mathbf{z}) \) together with the corresponding inverse samples \( \{\tilde{\mathbf{z}}\} \) from the real data, we optimize \( G \), \( D \), and \( F \) by adversarial learning to push \( \{\tilde{\mathbf{z}}\} \) closer to the sampling results of \( p(\mathbf{z}) \). Then we can obtain the final loss as:

\[
\min_{G,F} \max_{D} \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z})} \mathbb{E}_{\mathbf{x} \sim p_r(\mathbf{x})} [V(G, D)] + \lambda_{\text{re}} L_{\text{re}}(G, F) + \lambda_{\text{Gauss}} L_{\text{Gauss}}(F) \quad \tag{11}
\]

where \( L_{\text{Gauss}}(F) \) can be specified according to different distances or divergences and \( \lambda_{\text{re}}, \lambda_{\text{Gauss}} \) are weight parameters, which will be discussed in detail in the experiment section.

**Remarks for Conditional Generation.** IID-GAN can be extended to the conditioned case [Mirza and Osindero, 2014] when the label of the real data \( c \) is known. As for conditional IID-GAN, \( (\mathbf{z}, c) \) are the inputs of the generator \( G \) and the real \( \mathbf{x} \) is the input of \( F \) for classification. The Gaussian loss is used
to maintain the independence condition and learn the diversity of hidden features (e.g., the thickness for MNIST).

Remarks for the Disentanglement View. Many previous studies [Higgins et al., 2017; Kim and Mnih, 2018] learn the unsupervised disentanglement representation with the assumption of independent factors, i.e., \( q(z) = \prod_{j=1}^{M} q(z_j) \). Our work presents a new viewpoint with an \( M \)-D Gaussian guarantee. When \( q(z) \) approximates an \( M \)-D standard Gaussian, it is obvious that \( z \) is independent for different dimensions. However, varying \( z_i \) can not disentangle different modes as shown in the second column of Fig. 3. We observe that representing the data in polar coordinates and varying the polar angle and diameter can achieve better performance for unsupervised disentanglement representations.

**IID testing for Gaussian.** The regularization assumes a non-standard Gaussian and enforces the consistency of two Gaussians, which raises the doubt: Is the consistency useful for IID? We apply mathematical statistics here to test the Gaussian IID property. Specifically, we adopt the QQ-plot, Shapiro–Wilk test (SW) [Shapiro and Wilk, 1965], Kolmogorov-Smirnov test (KS) [Gosset, 1987] to show whether the samples are IID sampled from standard Gaussian. The results are shown in Fig. 5 and Table 2.

5 Experiments and Discussion

Experiments are conducted on a single GeForce RTX 3090. Synthetic data results are performed on GeForce RTX 2080Ti.

### 5.1 Experiments on Synthetic Datasets

Since the distribution is known for synthetic data, mode collapse can be directly measured. In line with [Metz et al., 2017], we simulate two synthetic datasets: i) Ring dataset. It consists of a mixture of 8 2-D Gaussians \( p(z) \) with mean \( \{ (2 \cos (i \pi / 4), 2 \cos (i \pi / 4)) \} \) and standard deviation 0.001. 12.5K samples are simulated from each Gaussian (100K samples in total). ii) Grid dataset. It consists of a mixture of 25 2-D isotropic Gaussians i.e. \( p(z) \) with mean \( \{ (2i, 2j) \} \) and standard deviation 0.0025. 4K samples are simulated from each Gaussian (100K samples in total).

### Metrics and network architecture

Following [Metz et al., 2017; Elfeki et al., 2019], we adopt the number of covered modes, generation quality\(^2\) and reverse KL divergence as the evaluation metrics. Since in the experiment, each mode shares the same number of real samples, one can calculate the reverse KL divergence between the generated distribution and the real one [Nguyen et al., 2017]. Note the reverse KL divergence is not strictly defined, as \( \sum_{i=1}^{m} \beta_i < 1 \) (there exist invalid generated points), thus it allows being negative. We adopt network architectures consisting of three linear layers with hidden dimensions of 100, 200, and 100, along with ReLU activation, for both the generator and discriminator.

#### Quantitative results.

The comparison involve vanilla GAN [Goodfellow et al., 2014], BiGAN [Donahue et al., 2017], Unrolled GAN [Metz et al., 2017] and VEEGAN [Srivastava et al., 2017] on Ring and Grid datasets in Table 1. IID-GAN variants are the only methods covering all modes on both datasets and IIGD-GAN (M-D) achieves the best quality and RKL performance compared to other peer methods.

#### Visualization with Gaussian inverse samples

Models involving inverse mappings (i.e. encoders) allow the visualization for the inverses of real samples. Fig. 3 present the visualization results including the inverse from real data, source Gaussian samples and generated results of VAE [Kingma and Welling, 2014], VEEGAN, BiGAN and IID-GAN variants. As shown in the left four columns, VAE can cause the overlap of the inverse samples \( z \), which may lead to bad generations (white points). BiGAN and VEEGAN learn the relations between \( z \) and \( x \) rather than the relation among the samples \( \{ G^{-1}(x) \} \), which leads to the failure of constructing 2D inverse Gaussian samples as shown in the first column. For IID-GAN, as shown in the fifth column for IID-GAN (M-D), the inverse samples are very similar to the Gaussian samples, which show the effectiveness of the regularization.

| Models          | 2D-Ring | 2D-Grid |
|-----------------|---------|---------|
| GAN             | 3.6 ± 0.5 | 96.6 ± 0.92 | 98.8 ± 0.06 | 0.75 ± 0.25 |
| BiGAN           | 6.8 ± 1.0 | 98.6 ± 0.95 | 98.9 ± 0.29 | 0.28 ± 0.30 |
| Unrolled GAN    | 6.4 ± 2.2 | 98.6 ± 0.50 | 98.7 ± 0.6 | 1.27 ± 0.17 |
| VEEGAN          | 5.4 ± 1.2 | 98.4 ± 0.67 | 98.0 ± 0.59 | 0.41 ± 0.10 |

Table 1: Results for Ring and Grid synthetic datasets.

\(^2\)We follow [De Meulemeester et al., 2021]: if the generated data is within 3 times std of the Gaussian, consider it a valid generation (otherwise bad). The resulting ratio is used as the generation quality.
Figure 3: Example for 2-D source (x) to 2-D target (x) generation and inverse adopting 8-mode Ring dataset as the real training set. Left 4 columns are the results of VAE, BiGAN, VEEGAN, while the right 4 columns are the results of IID-GAN under different Gaussian consistency losses as detailed in Sec. 4.3 after training 24K batches. For each half part, column 1 and column 5 show the inverse of the real target data, column 2 and column 6 show the sampled z from Gaussian in source space. Source points are in nine colors (8 ‘modes’ + 1 ‘bad’) according to their generation’s mode in the target space. The pie charts show the ratio of valid generation points in different modes.

Figure 4: Generation results on Ring and Grid datasets. Blue: generated samples; Red: ground truth distribution.

Figure 5: QQ-plot of standard 2-D Gaussian over two dimensions. The closer to the diagonal, the closer to the Gaussian distribution.

Figure 6: Generation quality and KL divergence (for diversity) calculated through the generations’ classification results.

significantly outperforms other peer methods in terms of mode coverage and better fits the target distribution.

IID test for inverse samples. We support IID test for Gaussian distribution on Ring dataset in Table 2. Shapiro-Wilk (SW) and Kolmogorov-Smirnov (KS) Statics are calculated to verify IID property. In Fig. 5, QQ-plot are supported for IID test. Compared to other peer methods, IID-GAN’s plots are closer to the red diagonal, implying that the inverse samples of IID-GAN are closer to Gaussian.

Ablation study for Gaussian consistency loss. The right four columns of Fig. 3 compare the results of 1-D, M-D and without Gaussian consistency loss in Ring datasets. For 1-D loss’s inverse samples in source space, the yellow sample region in the red box leads to a tendency to mode collapse, while the generated modes in the M-D case are more uniformly scaled and more effective for solving mode collapse.

5.2 Experiments on Real-world Data

The experimented image datasets include MNIST [LeCun et al., 1998], Stacked MNIST [Metz et al., 2017], CIFAR-10 [Krizhevsky et al., 2009], STL-10 [Coates et al., 2011], LSUN [Yu et al., 2015] and CELEBA [Liu et al., 2015]. We adopt different architectures to evaluate our model, mainly following the previous studies [Radford et al., 2016; De Meulemeester et al., 2021; Miyato et al., 2018]. All the compared models are trained in 100K steps. Details about the network architectures are presented in Appendix B.2. We generally adopt p-norm distance as M-D Gaussian loss. Please refer to Appendix B.3 for more details.

Metrics. We use Inception Score (IS) [Salimans et al., 2017] and Fréchet Inception Distance (FID) [Heusel et al., 2017]. Following [Richardson and Weiss, 2018], JSD measures the distribution distance between generated images and real images through clustering, which generally capture the distribution of image modes, thus it is more diversity related. For easily distinguishable datasets like MNIST, we use the number of covered modes and reverse KL for evaluation.

Training stability. In Fig. 6, the generation quality and KL divergence are evaluated on MNIST. The generation quality
We evaluate IID-GAN with StyleGAN2-ada [Karras et al., 2020] backbone on 64×64 resized AFHQ, LSUN via FID and KID metrics for same training iterations, i.e. 10M images trained (measured in real images shown to the discriminator is defined as the ratio of high-confidence samples evaluated by a trained classifier. 1-D and M-D IID-GAN can lead to less mode imbalance (which is evaluated by KL divergence) and higher quality than VAE, while VEEGAN is unstable and often fails to successfully generate in case of bad initialization.  

**Quantitative results.** (i) Table 3 show the #Mode, KL divergence and FID results on Stacked MNIST. The KL divergence calculation is based on the classification results by the classifier proposed by [Ding et al., 2019]. The architecture is DCGAN-based in line with [Radford et al., 2016]. (ii) Table 4 shows the experimental results of CIFAR-10 [Krizhevsky et al., 2009] and STL-10 [Coates et al., 2011]. We adopt JSD, IS, FID as the metrics, including UnrolledGAN [Metz et al., 2017], VEEGAN [Srivastava et al., 2017], MADGAN [Ghosh et al., 2018], GDDP [Elfeki et al., 2019] and BuresGAN (De Meulmeester et al., 2021) for comparison. To fit with advanced GAN achievements, we adopt SNGAN [Miyato et al., 2018] as the backbone. (iii) We also evaluate on larger scale datasets e.g. CELEBA [Liu et al., 2015] and LSUN Church [Yu et al., 2015] with more recently proposed works pursuing diversity for comparison including Dist-GAN [Tran et al., 2018], MGGAN [Bang and Shim, 2021] and AMAT [Mangalam and Garg, 2021]. We adopt FID, JSD as the evaluation metrics. The results are presented in Table 5. (iv) We evaluate IID-GAN with StyleGAN2-ada [Karras et al., 2020] backbone on 64×64 resized AFHQ, LSUN via FID and KID metrics for same training iterations, i.e. 10M images trained (measured in real images shown to the discriminator.
Appendix

A Gaussian Loss with Different Divergence

In this paper, Gaussian Consistency loss involves three different forms to reduce mode collapse:

1) p-norm for the difference of mean and variance. To evaluate the divergence of two Gaussian distributions \( \mathcal{N}(z; 0, I) \) and \( \mathcal{N}(z; \tilde{\mu}, \tilde{\Sigma}) \), we first calculate the difference of the parameters of Gaussian with p-norm:

\[
L_{\text{Gau}} = ||\tilde{\mu}||_p + ||\tilde{\Sigma} - I||_p
\]  
(12)

2) Wasserstein distance. Given two M-D Gaussians \( p(z) \) and \( q(z) \), the 2-Wasserstein distance is: \( W_2(p(z), q(z)) = 0 \) if and only if \( \tilde{\mu} = 0 \) and \( \tilde{\Sigma} = I \).

3) KL divergence. Given M-D Gaussians \( p(z) \) and \( q(z) \), the KL divergence \( KL(p(z), q(z)) \) can be specified as:

\[
L_{\text{Gau}} = \frac{1}{2} \left\{ \log(\det(\tilde{\Sigma})) - M + \text{tr}(\tilde{\Sigma}^{-1}) + \tilde{\mu}^T \tilde{\Sigma}^{-1} \tilde{\mu} \right\}
\]  
(13)

Results of IID-GAN with different Gaussian losses evaluated on synthetic dataset are presented in Fig. 9.

![Generation quality](image1.png)  ![Inverse KL divergence](image2.png)

(a) Generation quality  (b) Inverse KL divergence

Figure 9: Comparison of Gaussian consistency loss on Ring data

B Experimental Details

B.1 Training Details for Synthetic Data

We set the loss weights as \( (\lambda_{rec}, \lambda_{\text{Gau}}) = (3, 10) \). Training batch size is set as 100 and we conduct 300 epochs for training. The metrics are calculated over 5 repetitions of evaluation.

B.2 Network Architectures for Real Data

The architectures for CIFAR-10 are presented in Table 7 and 8. The architectures for CELEBA and LSUN are presented in Table 9 and 10. The inverse mapping primarily follows the design of discriminator, with only the dimension of the output layer changed to match that of the latent space.

B.3 Training Details for Real Data

MNIST. We set the loss weights as \( (\lambda_{rec}, \lambda_{\text{Gau}}) = (0.5, 0.1) \). Following [Dieng et al., 2019], we use the 10-category classifier to divide the generated images into 11 categories. If the highest probability of the generated image’s prediction is smaller than 0.75, we treat it as a bad generation and classify it to the bad class. Otherwise, its label is determined according to the highest probability. Latent space dimension is set as 2. We train the models for 300 epochs.

StackedMNIST. The StackedMNIST covers 1,000 known modes, as constructed by stacking three randomly sampled MNIST images along the RGB channels in line with the practice in [Metz et al., 2017]. We follow [Srivastava et al., 2017] to evaluate the number of covered modes and divergence between the real and generation distributions. The weights are set as \( (\lambda_{rec}, \lambda_{\text{Gau}}) = (3, 3) \). Latent space dimension is set as 100. We train the models for 50 epochs.

CIFAR-10, STL-10, CELEBA and LSUN. All models are trained for 100K steps (mini-batches). We set \( (\lambda_{rec}, \lambda_{\text{Gau}}) = (0.01, 0.1) \). Latent space dimension is set as 128 and training batch size is set as 128. The metrics are calculated over 3 repetitions of evaluation. Fig. 10 presents the generation results of IID-GAN on CELEBA and LSUN Church datasets.

![CELEBA](image3.png)  ![LSUN Church](image4.png)

(a) CELEBA  (b) LSUN Church

Figure 10: Generation results on CELEBA and LSUN Church for unconditional IID-GAN.

Contribution Statement

Yang Li and Liangliang Shi contribute equally to this work. Junchi Yan is the correspondence author. The work was partly supported by NSFC (62222607, U19B2035).
[Nguyen et al., 2017] T. Nguyen, T. Le, H. Vu, and D. Phung. Dual discriminator generative adversarial nets. In NIPS, 2017.

[Peyré et al., 2019] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. Foundations and Trends® in Machine Learning, 11(5-6):355–607, 2019.

[Radford et al., 2016] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. ICLR, 2016.

[Rezende and Mohamed, 2015] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. In ICML, 2015.

[Richardson and Weiss, 2018] Eitan Richardson and Yair Weiss. On gans and gmms. arXiv preprint arXiv:1805.12462, 2018.

[Salimans et al., 2017] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. In NIPS, 2017.

[Shapiro and Wilk, 1965] Samuel Sanford Shapiro and Martin B Wilk. An analysis of variance test for normality (complete samples). Biometrika, 52(3/4):591–611, 1965.

[Srivastava et al., 2017] Akash Srivastava, Lazar Valkov, Chris Russell, Michael Gutmann, and Charles Sutton. Vegan: Reducing mode collapse in gans using implicit variational learning. In NIPS, 2017.

[Tran et al., 2018] Ngoc-Trung Tran, Tuan-Anh Bui, and Ngai-Man Cheung. Dist-gan: An improved gan using distance constraints. In ECCV, 2018.

[Yang et al., 2019] Dingdong Yang, Seunghoon Hong, Yunseok Jang, Tianchen Zhao, and Honglak Lee. Diversity-sensitive conditional generative adversarial networks. In ICLR, 2019.

[Yu et al., 2015] Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. arXiv preprint arXiv:1506.03365, 2015.

[Yu et al., 2019] Xianwen Yu, Xiaoning Zhang, Yang Cao, and Min Xia. Vaegan: A collaborative filtering framework based on adversarial variational autoencoders. In IJCAI, pages 4206–4212, 2019.