Implicit Discriminator in Variational Autoencoder

Prateek Munjal*
Indian Institute of Technology Ropar
India
2017csm1009@iitrpr.ac.in

Akanksha Paul*
Indian Institute of Technology Ropar
India
akanksha.paul@iitrpr.ac.in

Narayanan C. Krishnan
Indian Institute of Technology Ropar
India
ckn@iitrpr.ac.in

Abstract—**Recently generative models have focused on combining the advantages of variational autoencoders (VAE) and generative adversarial networks (GAN) for good reconstruction and generative abilities. In this work we introduce a novel hybrid architecture, Implicit Discriminator in Variational Autoencoder (IDVAE), that combines a VAE and a GAN, which does not need an explicit discriminator network. The fundamental premise of the IDVAE architecture is that the encoder of a VAE and the discriminator of a GAN utilize common features and therefore can be trained as a shared network, while the decoder of the VAE and the generator of the GAN can be combined to learn a single network. This results in a simple two-tier architecture that has the properties of both a VAE and a GAN. The qualitative and quantitative experiments on real-world benchmark datasets demonstrate that IDVAE performs better than the state of the art hybrid approaches. We experimentally validate that IDVAE can be easily extended to work in a conditional setting and demonstrate its performance on complex datasets.**

Index Terms—**Variational Autoencoders, Generative Adversarial Networks**

I. Introduction

Deep Variational Autoencoders(VAE [15]) and Generative Adversarial Networks(GAN [12]) are two recently used approaches in the generative modeling world. VAE is more stable in training but generates blurry samples. While GAN has the appealing property of generating realistic images; training a GAN is well known to be challenging leading to problems such as mode collapse.

Several recent approaches have proposed hybrid models of autoencoder and adversarial networks with a joint objective of achieving stable training like VAE and inferencing ability like GAN. In order to introduce the adversarial loss component in the objective functions most of the recent hybrid approaches include an adversary network that results in a three-tier architecture i.e an encoder, a decoder, and an adversary network. We hypothesize that the encoder and discriminator networks can share common layers encoder itself can be reused as a discriminator, thereby assuming an overlap in the knowledge learned by the encoder and the discriminator network.

The key idea behind our approach is that we would like the discriminator to provide the useful gradients to the generator if it misses any mode in the true data distribution. The traditional GAN learning does not explicitly encourage such a property in the discriminator and therefore, we suspect that it is vulnerable
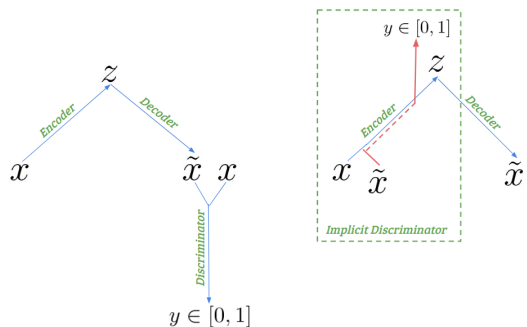
Fig. 1: Flow diagram of traditional hybrid approaches(left) and our proposed approach(right). We introduce the adversarial loss by collapsing the encoder into the discriminator, which we term as Implicit Discriminator. The output of discriminator is denoted by $y \in [0, 1]$ where 0 and 1 represent fake and real respectively.

to the issue of mode collapse. Further, we note that the traditional L2 loss used for learning the encoder can be seen as minimizing the forward-KL divergence. The forward-KL divergence comes with the mode inclusive property (i.e it never misses a mode in the true data distribution). Therefore, to make the discriminator aware of all the modes in the true data we propose to share the forward-KL information with the discriminator by explicitly sharing the parameters between encoder and the discriminator network. We restrict the sharing of the parameters between the encoder and discriminator networks until the penultimate layer to facilitate the modelling of the different outputs.

Figure 1 illustrates our proposed two-tier architecture and contrasts it against the traditional hybrid (V)AE/GAN approaches. We propose an adversary free two-tier architecture an encoder and decoder network that has the capabilities of both a discriminator and a generator. In the proposed model the decoder network collapses into the generator, while the encoder network is merged with the discriminator resulting in a two tier architecture. We term our two tier architecture as Implicit Discriminator in Variational Autoencoder (IDVAE). VAE-GAN is a special case of IDVAE where there is no sharing of parameters between the encoder and the discriminator.

In this work we show that our proposed simpler hybrid VAE/GAN model, IDVAE outperforms the prior approaches

in terms of visual fidelity measured in terms of FID score. We also show that our model sustains both reconstruction and generation ability without training an unnecessary adversary network that would result in learning more parameters. Overall the main contributions of the work are as follows :

- We introduce a novel two-tier architecture, IDVAE, which sustains the abilities of both reconstruction (VAE) and generation (GAN) without learning a separate discriminator or a generator.
- We present a training schedule that facilitates the encoder to act as as implicit discriminator while maintaining the tight coupling between encoder and decoder network.
- Empirical evaluations of IDVAE performed on benchmark datasets show that IDVAE achieves better generative ability than prior approaches. We also show that Fréchet Inception Distance, a common measure to evaluate the quality of the generations has inconsistent outputs and thereby propose an ensemble of experts for conducting quantitative evaluation.

II. RELATED WORK

Variational Autoencoder (VAE) introduced by Kingma et al. [15] minimizes the KL divergence between the real distribution (P_x) and the generated distribution (P_g) through the variational bound. Detailed analysis of VAE by Doersch [9] shows that VAE works well in practice and is considered to model the true data distribution quite well but often generates poor quality samples the images produced by the decoder are blurred. On the other hand GAN's [12] generate samples that are visually more realistic through an adversarial game play between the generator and the discriminator. However, GAN's suffer from problems like instability during training and mode collapse. Recently Bang and Shim [4] proposed RFGAN that uses pre-trained encoder features (representative features) to regularize the training of the discriminator to alleviate the problem of mode collapse. Similarly, MR-GAN [6] also proposes to use autoencoder features as regularizer in GAN training. Inspired from these architectures, we propose IDVAE that exploits the complementary properties of forward KL and reverse KL to capture the data distribution. While these approaches make use of a pre-trained encoder, our approach jointly and simultaneously trains a VAE and GAN achieving both the reconstruction and generation capabilities. There has also been some efforts towards utilizing the advantages of VAE for training GANs. Larsen [17] proposed VAE-GAN that collapses the decoder of the VAE into the generator of the GAN. VAE-GAN achieves sharp generations using a similarity metric learned by the intermediate representations of an explicit adversary. VAE-GAN requires an explicit discriminator, while our proposed approach overcomes this necessity by converting the encoder of the network into a discriminator. ALI [11] and BiGAN [10] also propose to use three networks: the encoder, the decoder and the adversary. Unlike IDVAE, both the ALI and the BiGAN discriminator differentiates between samples from the joint distribution of observed data and latent codes. However, the reported reconstructions are of poor quality [21]. Akin to BiGAN discriminator setting, the AVB [21] model uses an additional discriminator to facilitate learning without explicitly assuming any form for posterior distribution. However, the samples generated by AVB for the CelebA dataset are observed to be blurry [3]. In contrast, the simpler IDVAE model is able generate higher quality samples with lesser parameters.

Li [19] propose ALICE that improves upon ALI by alleviating certain undesirable solutions (saddle points). Unlike a two tier approach of IDVAE, ALICE requires three networks and proposes to regularize the objective with cycle loss (an upper bound for conditional entropy). While in IDVAE, there is an implicit regularization on the discriminator by sharing its parameters with the encoder. The AS-VAE model of Pu [23] focuses on both reverse and forward-KL between the encoder and decoder joint distributions with an objective to maximize the marginal likelihood of observations and latent codes. AS-VAE also needs two adversaries to circumvent the need of assuming an explicit form for the true intractable distribution (eqn 8 and 9 in [21]). IDVAE also focuses on forward-KL and reverse-KL but in a very novel way by sharing the parameters of the encoder (forward-KL) and discriminator (reverse-KL) resulting in a simpler model. α-GAN [24] fuses VAE and GAN exploiting the density ratio trick by constructing two additional discriminators for measuring the divergence between the reconstructions and the true data points, and the latent representations and the latent prior. The first discriminator minimizes the reverse-KL divergence, and the reconstruction error term minimizes the forward-KL divergence to discourage mode collapse. Training α-GAN is difficult as it requires learning a large set of parameters (for the 4 networks). In contrast to previous approaches, Ulyanov [28] propose a two tier adversary free approach, AGE, where the encoder network is responsible for the adversarial signal. While the architectures of AGE and IDVAE appear to be similar, there are some fundamental differences in the process of learning the discriminator. The AGE discriminator compares (via divergence) the encoded real and fake distributions against a fixed reference distribution (typically, a prior in latent space). Whereas the IDVAE discriminator directly compares the real and fake data using a simple cross entropy loss, where both the reconstructions and randomly generated samples are treated as fake examples. We empirically show that IDVAE learns better as its discriminator relies on reconstructed samples as well. Importance of reconstructed samples in adversarial learning is supported in literature [7].

III. METHODOLOGY

Notations Let $\mathbf{x}$ be the data point in the input space $\mathcal{X}$ and $\mathbf{z}$ be the code in the latent space $\mathcal{Z}$. The output of the encoder and the discriminator network for an input $\mathbf{x}$ is represented as $\text{Enc}(\mathbf{x})$ and $\text{Dis}(\mathbf{x})$ respectively. Similarly the output of decoder network $\tilde{\mathbf{x}}$ for a latent code $\mathbf{z}$ is denoted by $\text{Dec}(\mathbf{z})$. The output at the lth layer of the encoder network for an input $\mathbf{x}$ is denoted as $\text{Enc}_l(\mathbf{x})$. This is same as the output at the lth layer of the discriminator network for an
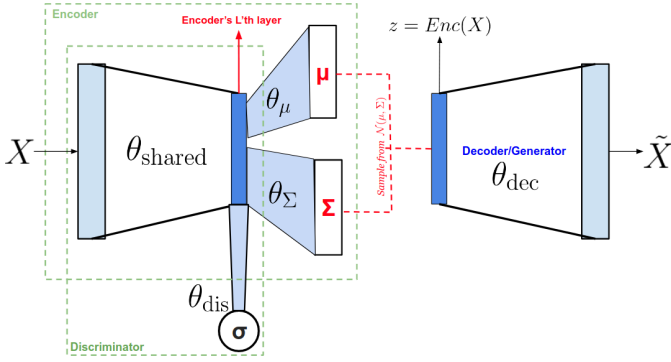
Fig. 2: The proposed architecture for IDVAE. The parameters θ_μ and θ_Σ denote a single fully connected layer learning the encoder specific parameters, θ_{enc}. θ_{dis} also represents a single fully connected layer which denotes discriminator specific parameters. Similarly θ_{shared} denotes the shared parameters between the encoder and the discriminator whereas θ_{dec} denotes the decoder/generator specific parameters.

input $\mathbf{x}$ which is denoted as $\text{Dis}_l(\mathbf{x})$. $\text{Enc}_l(\mathbf{x})$ and $\text{Dis}_l(\mathbf{x})$ are used interchangeably depending on the context. In reference to Figure 2, we denote the encoder specific parameters by θ_{enc} where $\theta_{enc} = \{\theta_\mu, \theta_\Sigma\}$.

We start with some preliminaries on VAE [15] and GAN [12] before describing our proposed model, which combines both of them.

A. Variational Autoencoder

A VAE comprises of learning two networks, namely, the encoder and the decoder network. In contrast to traditional autoencoders, VAE views the encoder and the decoder networks as probabilistic functions. The encoder learns a conditional distribution on the latent code $\mathbf{z}$ conditioned on the input $\mathbf{x}$. Similarly decoder learns a distribution on $\tilde{\mathbf{x}}$ conditioned on the latent code $\mathbf{z}$.

$$\mathbf{z} \sim \text{Enc}(\mathbf{x}) = q(\mathbf{z}|\mathbf{x}) \qquad (1)$$
$$\tilde{\mathbf{x}} \sim \text{Dec}(\mathbf{z}) = p(\mathbf{x}|\mathbf{z}) \qquad (2)$$

Vanilla VAE jointly trains over the encoder and the decoder network parameters by minimizing negative log-likelihood (reconstruction term) and divergence between prior and learned distribution in latent space $\mathcal{Z}$. The prior, $p(\mathbf{z})$, over the latent space is typically assumed to be a unit Normal distribution, i.e. $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Thus training a VAE would mean minimizing the following loss:

$$\mathcal{L}_{\text{VAE}} = \mathcal{L}_{\text{recons}} + \mathcal{L}_{\text{prior}} \qquad (3)$$

where,

$$\mathcal{L}_{\text{recons}} = -\mathbb{E}_{q(\mathbf{z}|\mathbf{x})}[\log p(\mathbf{x}|\mathbf{z})] \qquad (4)$$
$$\mathcal{L}_{\text{prior}} = KL(q(\mathbf{z}|\mathbf{x})\|p(\mathbf{z})) \qquad (5)$$

and $KL(q(\mathbf{z}|\mathbf{x})\|p(\mathbf{z}))$ is the Kullback-Leibler divergence between the distributions $q(\mathbf{z}|\mathbf{x})$ and $p(\mathbf{z})$.

B. Generative Adversarial Network

A GAN consists of two networks, namely, the discriminator and the generator that are learned through an adversarial game play. The generator network maps a point $\mathbf{z}$ in some arbitrarily low dimensional latent space $\mathcal{Z}$ to a point in a high dimensional data space $\mathcal{X}$. We denote $\text{Gen}(\mathbf{z})$ as the output of the generator network when $\mathbf{z}$ is the input. In a similar vein, the discriminator network maps a data point $\mathbf{x}$ in the data space to a probability value $y \in [0, 1]$. The objective of the discriminator is to assign the probability $y = \text{Dis}(\mathbf{x})$ that $\mathbf{x}$ is a sample from true distribution and the probability $1 - y$ that $\mathbf{x}$ is a generated sample i.e. $\mathbf{x} = \text{Gen}(\mathbf{z})$, with $\mathbf{z} \sim p(\mathbf{z})$. Thus, in this adversarial game play, the objective of the generator is to synthesize samples that can fool the discriminator learning the true data distribution, while the goal of the discriminator is to recognize the samples coming out of the generated(fake) distribution and the true distribution. Adversarial game play between the discriminator and the generator is formally defined by the GAN loss as

$$\mathcal{L}_{GAN} = \log(\text{Dis}(\mathbf{x})) + \log(1 - \text{Dis}(\text{Gen}(\mathbf{z}))) \qquad (6)$$

We want to maximize the binary cross entropy loss with respect to the discriminator (D) while minimizing it for the generator (G). Thus, the minimax objective is defined as

$$\min_G \max_D \mathcal{L}_{GAN} \qquad (7)$$

C. IDVAE

Our proposed approach, **I**mplicit **D**iscriminator in **V**ariational **A**uto**E**ncoder (IDVAE), exploits the properties of both VAE and GAN. IDVAE sustains the stable training properties of VAE while generating samples of quality approaching GAN. We borrow the encoder and decoder networks from VAE with slight modifications. In particular, we collapse the VAE decoder network into the generator of the GAN and the VAE encoder network partially into the discriminator of the GAN.

We partially collapse the encoder into the discriminator following the assumption that there exists an overlap in the knowledge of encoder and discriminator network. As the encoder's objective is to learn representational features, while the discriminator's objective is to learn discriminative features, we restrict the weight sharing to the penultimate layer (say some l^{th} layer of encoder represented as Enc_l) in encoder of the VAE. Further to encourage the encoder to learn the features of discriminator we add a single fully connected layer from Enc_l to a single sigmoid node that acts as the discriminator's output. Figure 2 illustrates the proposed IDVAE network architecture.

Thus, in our model we have four sets of parameters that need to be learned, namely; θ_{dec} - the shared parameters between the decoder and the generator, θ_{shared} - the parameters shared between the encoder and the discriminator, θ_{enc} - the encoder specific parameters of the VAE, and θ_{dis} - the discriminator specific parameters of the GAN. These are updated based on the loss incurred by each of the individual networks.

The loss incurred by the encoder is used to update both θ_{shared} and θ_{enc}. The encoder loss in the IDVAE consists of two components similar to a standard VAE. The first component is the reconstruction loss - $\mathcal{L}_{\text{recons}}$ and the second component is the prior discrepancy loss - $\mathcal{L}_{\text{prior}}$. It is well known that minimizing the forward-KL divergence $\text{KL}(P_{data}\|P_{model})$ achieves mode coverage for generative models. Thus, we minimize forward-KL divergence by minimizing $\mathcal{L}_{\text{recons}}$ for helping IDVAE to learn the different modes in the data. The shared parameters between the encoder and the discriminator encode the forward-KL divergence information. Thus using $\mathcal{L}_{\text{recons}}$ in the encoder reduces the extent of mode collapse as the gradients from the discriminator to the generator implicitly contain the information about multiple modes. Thus the overall encoder loss ($\mathcal{L}_{enc}$) is defined as follows

$$\mathcal{L}_{\text{enc}} = \alpha \mathcal{L}_{\text{recons}} + \beta \mathcal{L}_{\text{prior}} \qquad (8)$$

where α and β are hyper parameters controlling the contribution of each of the loss terms.

It has been shown that GAN [12] achieves sharper images by minimizing the reverse-KL divergence. Thus, we use the implicit adversary (encoder as a shared discriminator) of IDVAE as a way to propagate reverse-KL divergence information. The discriminator loss is used to update both θ_{shared} and θ_{dis}. The generated (fake) examples that are presented to the discriminator of IDVAE are the output of the decoder when viewed as a generator $\text{Dec}(\mathbf{z})$, where $\mathbf{z} \sim p(\mathbf{z})$. In addition to this, we also present the synthesized sample through reconstruction, $\text{Dec}(\text{Enc}(\mathbf{x}))$, for an input $\mathbf{x}$. As $\text{Dec}(\text{Enc}(\mathbf{x}))$ is more likely to be similar to $\mathbf{x}$ than $\text{Dec}(\mathbf{z})$, for an arbitrary $\mathbf{z} \sim p(\mathbf{z})$, we hypothesize that the discriminator loss corresponding to $\text{Dec}(\text{Enc}(\mathbf{x}))$ encourages the generator to learn the properties of the decoder. Similarly the discriminator loss corresponding to $\text{Dec}(\mathbf{z})$ encourages the decoder to learn the properties of the generator be able to generate realistic examples from the prior distribution $p(\mathbf{z})$. Therefore using both the terms, $\text{Dec}(\mathbf{z})$ and $\text{Dec}(\text{Enc}(\mathbf{x}))$ encourages the model to learn a blend of both the generator and the decoder. Intuitively in equation 9 to maintain the ratio of real and fake samples shown to the discriminator the loss terms for the fake samples should be scaled by a factor of 0.5 or the real term $\log(\text{Dis}(\mathbf{x}))$ by 2. We observed no significant change in the performance of IDVAE when these factors are dropped, thereby giving rise to the following loss function

$$\mathcal{L}_{\text{dis}} = -\big[\log(\text{Dis}(\mathbf{x})) + \log(1 - \text{Dis}(\text{Dec}(\mathbf{z})))$$
$$+ \log(1 - \text{Dis}(\text{Dec}(\text{Enc}(\mathbf{x}))))\big] \quad (9)$$

As we have collapsed the decoder of the vanilla VAE into the generator, the loss incurred by both the decoder and generator is used to update the shared parameters between the decoder and generator (θ_{dec}). The decoder/generator loss in IDVAE consists of two components. The first component ($\mathcal{L}_{\text{recons}}^{\text{dis}}$) is a learned similarity metric motivated by VAE-GAN [17]. Specifically, we learn a similarity metric($\mathcal{L}_{\text{recons}}^{\text{dis}}$) using an intermediate representation (l^{th} layer) of the discriminator

(equivalent to the lth layer of the encoder) by assuming a Gaussian observation model on $\text{Dis}_l(\tilde{\mathbf{x}})$ with mean $\text{Dis}_l(\mathbf{x})$ and unit covariance :

$$p(\text{Dis}_l(\tilde{\mathbf{x}})|\mathbf{z}) = \mathcal{N}(\text{Dis}_l(\tilde{\mathbf{x}})|\text{Dis}_l(\mathbf{x}), \mathbf{I}) \qquad (10)$$

where for a given sample $\mathbf{x}$, $\tilde{\mathbf{x}} = \text{Dec}(\mathbf{z})$ and $\mathbf{z} = \text{Enc}(\mathbf{x})$. $\mathcal{L}_{\text{recons}}^{\text{dis}}$ is defined as a Gaussian observation model:

$$\mathcal{L}_{\text{recons}}^{\text{dis}} = -\mathbb{E}_{q(\mathbf{z}|\mathbf{x})}[\log p(\text{Dis}_l(\mathbf{x})|\mathbf{z})] \qquad (11)$$

The second component is the adversarial loss which encourages the decoder to learn the properties of a generator. The adversarial loss, ($\mathcal{L}_{\text{GAN}}$), is defined as

$$\mathcal{L}_{\text{GAN}} = -log(\text{Dis}(\text{Dec}(\mathbf{z}'))) - log(\text{Dis}(\text{Dec}(\text{Enc}(\mathbf{x})))) \quad (12)$$

where $\mathbf{z}' \sim p(\mathbf{z})$.

Therefore the presence of both the reconstruction loss and the adversarial loss in objective function of decoder makes it learn a blend of the two models. The overall loss function for the decoder/generator ($\mathcal{L}_{\text{dec}}$) is defined as:

$$\mathcal{L}_{\text{dec}} = \omega \mathcal{L}_{\text{GAN}} + \lambda \mathcal{L}_{\text{recons}}^{\text{dis}} \qquad (13)$$

where ω and λ used in $\mathcal{L}_{\text{dec}}$ are hyper-parameters that are learned empirically.

D. Training Schedule

The lth shared layer between the encoder and the discriminator outputs representations for learning the parameters of the encoder's distribution and for discriminating between samples from the true distribution and generated samples simultaneously. We need to ensure that the shared weights (θ_{shared}) of the encoder and discriminator network gets the learning signal corresponding to both the encoder and discriminator objective function. In theory $\theta_{\text{shared}}, \theta_{\text{dis}},$ and θ_{enc} can be updated in a single step using the joint loss of both the encoder and discriminator. However, in practise we observed training using the joint loss to be challenging in terms of hyper-parameter fine tuning. Hence, we update the shared weights (θ_{shared}) in two iterations, once each with the encoder and discriminator losses respectively. In the first iteration θ_{shared} and θ_{dis} are updated while in the second iteration θ_{shared} and θ_{enc} are updated. Algorithm 1 presents the overview of the training procedure. Thus the parameters θ_{shared} learn the information of both reverse-KL (first iteration) and forward-KL (second iteration), which is leveraged by the decoder/generator in the third step of the algorithm. The parameters, θ_{shared}, can be updated in any arbitrary order, we empirically found that using first the discriminator loss helps in better learning. We present the qualitative results on the other two variants (using the joint loss, and updating θ_{shared} with respect to the encoder first followed by the discriminator) in the supplementary material.

IV. EXPERIMENTS

We investigate the proposed IDVAE architecture for the quality of both reconstructions and generations. We evaluate the performance of IDVAE on the following two real world benchmark datasets: i) **CIFAR10** [16], which contains

Algorithm 1 IDVAE Training Schedule

1: $P(z) \leftarrow \mathcal{N}(0, I)$
2: $\theta_{\text{shared}}, \theta_{\text{enc}}, \theta_{\text{dec}}, \theta_{\text{dis}} \leftarrow$ Initialize parameters
3: $X \leftarrow$ random mini batch from dataset
4: $Z \leftarrow \text{Enc}(X)$
5: $\tilde{X} \leftarrow \text{Dec}(Z)$
6: $Z' \leftarrow$ samples from prior $P(Z)$
7: $X' \leftarrow \text{Dec}(Z')$
8: **while** *not convergence* **do**
9: $\quad \theta_{\text{dis}}, \theta_{\text{shared}} \xleftarrow{+} -\nabla_{(\theta_{\text{dis}}, \theta_{\text{shared}})}(\mathcal{L}_{\text{dis}})$
10: $\quad \theta_{\text{enc}}, \theta_{\text{shared}} \xleftarrow{+} -\nabla_{(\theta_{\text{enc}}, \theta_{\text{shared}})}(\mathcal{L}_{\text{enc}})$
11: $\quad \theta_{\text{dec}} \xleftarrow{+} -\nabla_{\theta_{\text{dec}}}(\mathcal{L}_{\text{dec}})$
12: **end while**

Fig. 3: Comparing reconstruction loss (lower is better) among different generative models.

60k images of which 50k are used for training and the remaining 10k for testing. (ii) **CelebA** [20], which consists of 202,599 images. We use 1-162,770 images for training, 162,771-182,637 for validation and rest for testing. In our implementation pipeline we crop and scale the images to 64x64 for faster training. The details of encoder and decoder network architecture along with fine tuned hyper-parameters for each of the dataset are provided in the supplementary material. For generating instances over the different datasets we randomly sample **z** from the assumed prior distribution (on the latent space $\mathcal{Z}$) $\mathcal{N}(\mathbf{0}, \mathbf{I})$. We also conducted experiments on a synthetic 2D Gaussian dataset and the MNIST digits dataset. These details can be found in the supplementary material.

We compare the performance of IDVAE against approaches that have both generative and reconstruction abilities, namely; VAE [15], VAE-GAN [17], AGE [27], and α-GAN [24]. We use the pre-trained models available for AGE, while we train all the other models from scratch using the best hyper-parameters reported in the literature.

V. RESULTS AND DISCUSSION

We conduct both qualitative and quantitative comparison of IDVAE for generations and reconstructions against all the prior approaches.

A. Quantitative Analysis

The reconstruction quality is objectively quantified using the standard square loss $\mathcal{L}_{\text{recons}}$. We obtain an unbiased estimate of the loss using a large test set consisting of 10k samples for both the CelebA and the CIFAR10 datasets. Thus even a small improvement on such a large set is significant considering the complexities of the dataset. It is evident from the results presented in Figure 3 that for the reconstruction task IDVAE performs better than or is at par with VAE-GAN, AGE and α-GAN. However the lowest reconstruction error is obtained by VAE. This is understandable as there is no explicit penalty on the decoder for reconstructing unrealistic images. On the other hand both VAE-GAN and IDVAE strike a balance between the reconstruction loss and the generative ability. We also modify IDVAE to explicitly minimize $\mathcal{L}_{\text{recons}}$ in the decoder, which
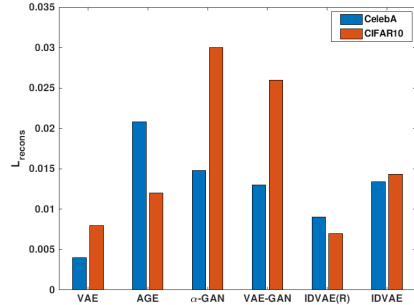
we term as IDVAE(R). The reconstruction loss of IDVAE(R) is the best among all the variants of VAE, However, this comes at the cost of quality of the generations.

There are two popular measures for qualitatively evaluating generative models, namely the Inception Score (IS) [25] and Fréchet Inception Distance (FID) [14]. It has been shown that IS closely follows human scoring of images synthesized by generative models for the CIFAR 10 dataset [16]. The IS uses the Inception v3 model pre-trained on ImageNet. The IS is a statistic on the Inception model's output when applied to the synthesized images. This statistic captures two desirable qualities of a generative model - the synthesized images should contain an object (the image is sharp and not blurry) that is reflected in a low entropy output of the Inception model [26] and; there must be diversity in the generations that is reflected in the high entropy output of the Inception model over the entire generated set. Barratt and Sharma [5] have recently shown that the IS suffers from suboptimalities and is an appropriate measure only for datasets that are trained on ImageNet. Thus, it is not advisable to measure the quality of generations on the CelebA dataset. The FID improves upon the IS by comparing the statistics of both the generated and true samples, instead of evaluating the generated samples in isolation. The FID is the Fréchet distance between two multivariate Gaussians estimated from the 2048-dimensional activations of the Inception-v3 pool3 layer for real and generated samples. Lower FID scores correspond to more similar real and synthetic samples.

Table I, presents the Fréchet distance scores computed using Inception-15[1] (FID_{15}), Inception-16[2] (FID_{16}), and ResNet [13]. As all the three experts have the same knowledge all the models are pre-trained on ImageNet [8] and the representations extracted from the intermediate layer have the same dimensions (2048), the relative performance of the models with respect to each expert can be compared. While the distance scores across the experts for the same model may be different, we expect the order of the goodness among the generative models to be preserved. However, as can be observed from

---

[1] weights used from 2015 year model [1]
[2] weights used from 2016 year model [2]

Model	FID_{15}	FID_{16}	FRD
CIFAR10			
True Data	3.16±0.06	7.41±0.82	26.17±1.44
IDVAE	23.48±0.15	28.15±0.39	**105.45±0.79**
IDVAE(R)	43.38±0.15	49.9±0.85	191.32±5.88
VAE-GAN	27.04±0.12	33.12±0.73	139.95±2.71
VAE	85.74±0.3	130.38±3.47	626.67±8.61
AGE	32.19±0.3	29.3±0.54	122.43±2.61
α-GAN	**20.61±0.12**	27.87±0.7	121.88 ± 3.09
CelebA			
True Data	1.58±0.02	2.67±0.15	5.77±0.35
IDVAE	**8.53±0.12**	9.52±0.72	**34.47±2.41**
IDVAE(R)	14.81±0.17	16.71±0.66	70.99±0.91
VAE-GAN	9.52±0.06	10.5±0.9	38.32±1.69
VAE	35.27±0.04	55.44±0.87	150.02±1.41
AGE	12.74±0.14	15.27±0.36	82.45±0.97
α-GAN	10.38±0.2	13.89±1.58	55.44±7.97

TABLE I: Comparing Frechet Distance (lower is better) among different generative models. $FID_{15,16}$: Inception Model with 2015, 2016 year weights respectively and FRD ResNet model.
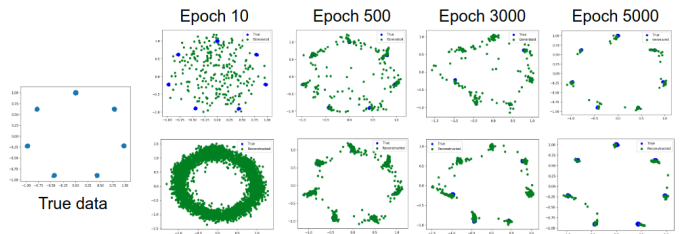

Fig. 4: Experiments on the 7 mode Gaussian mixture dataset. The leftmost image denotes the true data distribution. The true modes are colored in blue whereas points in green color are either reconstructed or generated data points. Images in top and bottom row show generated samples and reconstructed data points respectively.

Table I on CIFAR10 dataset, VAE-GAN appears to perform better than the AGE based on FID_{15}, while the trend reverses when comparing based on both FID_{16} and FRD. Therefore, our results suggest that a generative model should be compared across a battery of experts rather than in isolation. IDVAE performs better than all the other approaches on the CelebA dataset. The result is statistically significant on both FID_{15} and FRD scores. On the other hand both IDVAE and α-GAN result in the best performance on the CIFAR10 dataset. There is no significant difference between IDVAE and α-GAN with each performing better than the other only according to a single measure. However, from Figure 3, it is quite apparent that α-GAN focuses less on reconstructions whereas in IDVAE we do not observe such a bias. These results support our hypothesis that the encoder and discriminator can be a shared network. IDVAE is able to perform at par or sometimes better than VAE-GAN and α-GAN that require a separate encoder/discriminator network. This is further verified through our qualitative results. We also observe a dip in the Fréchet distance for the IDVAE(R) model in comparison to IDVAE. As the decoder/generator of IDVAE(R) focuses on reconstructions in the image space we observe a drop in the Fréchet distance at the cost of a better reconstruction loss. Therefore IDVAE(R) model has the potential to fit within the required thresholds by tuning the hyper parameters ω, λ and γ.

B. Qualitative Analysis

Mode collapse is the problem faced in GAN [12] when the generator stops exploring, thereby limiting the diversity in its generations. As the gradients coming from discriminator of IDVAE have the information of both the representative (encoder) and the discriminative (discriminator) features, we hypothesize that by using the encoder as a discriminator, the generator gets better learning signal than it would have received from an independent discriminator. We perform the mode collapse test using a toy 2D dataset of 7 mode Gaussian mixture to qualitatively investigate both the reconstructions (Figure 4) and the generations (Figure 4) of IDVAE. Our objective behind the mode collapse test is to show that IDVAE is free from mode collapse which is apparent from the above mentioned figures. In fact, it is effectively able to capture all the modes. Thus the empirical result further supports our hypothesis that **encoder can act as a implicit discriminator**. We present the qualitative results obtained from the different models in the Table II on the CIFAR10 and CelebA datasets. It is quite evident from the images that VAE results in blurry reconstructions while the rest of the approaches output sharp images, which is due to presence of adversarial loss. On both CelebA and CIFAR10 datasets, we observe IDVAE and IDVAE(R) performing on par with VAE-GAN and α-GAN, while significantly outperforming VAE in terms of sharpness of the images. The images generated by the different models are also presented in Table II. The undesirable blurriness property in VAE is apparent on the CIFAR10 dataset while the performance of IDVAE is on par with both α-GAN and VAE-GAN. The images generated by IDVAE trained on the CelebA dataset appears to capture a large diversity in background when compared to α-GAN and VAE-GAN. We observe both IDVAE(R) and α-GAN tend to focus more on faces than the background in these images.

C. Conditional IDVAE

We extend IDVAE to a conditional setting where our objective is to learn a generator/decoder whose output is controlled by some conditional information, y. We term this variant as **C**onditional-IDVAE (C-IDVAE). We qualitatively analyze the C-IDVAE model using MNIST [18] and CelebA datasets. We follow the recent work of Perarnau [22] to provide the conditional information to the generator at the input layer, while for the discriminator this is provided after the first convolution layer. To the best of our knowledge the encoder of VAE is never made aware of the conditional information but as we have our encoder acting as a discriminator we add this conditional information after the first convolution layer.

Fig. 5: Conditional-IDVAE on celebA and MNIST datasets

We use the one hot encoding of the MNIST class labels as the conditional information to evaluate C-IDVAE. Figure 5 illustrates the samples generated by C-IDVAE, where each row illustrates the images generated by conditioning on a unique label. We observe a large diversity in the generations in each row implying the diverse generative ability of the conditional decoder/generator.

Following the work of Perarnau [22] we use 13 attributes that have clear visual impact out of the total 40 attributes as conditioning information while training C-IDVAE on the CelebA dataset. Figure 5 presents the images generated by C-IDVAE for different conditioning information. Each row in the figure represents the images generated by C-IDVAE with the conditioning information provided in the top row, and the original image that is modified in the first column. It can be observed that the changes in each of the generations with respect to the original image are a result of the model imagining the original image on different attributes. For example, consider third and fifth rows in the Figure 5, where the original image is a female face. The generations in the columns 2, 5, and 6 that are conditioned on the male attribute actually contain a face that resembles a male. Similarly in the last column the model does reasonably well in adding eyeglasses to all the generated images. Thus considering the generative ability matching with the human imagination and the complexity of the real world CelebA dataset, C-IDVAE shows the potential to model the complex distributions.

VI. SUMMARY AND FUTURE WORK

In this work, IDVAE, a novel hybrid of the variational autoencoder and the generative adversarial network is introduced. IDVAE shares a common decoder and generator network, and partially shares the encoder and the discriminator network. The experiments on real-world benchmark datasets demonstrates that IDVAE (and its variant IDVAE(R)) performs on par and sometimes better than the state of the art hybrid approaches. We also show that IDVAE can be easily extended to work in a conditional setting, and experimentally demonstrate its performance on complex datasets. Further, our results present inadequacies of the Fréchet Inception Distance and suggests an ensemble of experts for evaluating the quality of the generations. This can be further explored to derive a measure that does not require a model that is pre-trained on data from a different domain as that of the training samples.

REFERENCES

[1] http://download.tensorflow.org/models/image/imagenet/inception-2015-12-05.tgz.
[2] http://download.tensorflow.org/models/inception$_v$32016$_0$8$_2$8.$tar.gz$.
[3] Official avb github repository. https://github.com/LMescheder/AdversarialVariationalBay
[4] D. Bang and H. Shim. Improved training of generative adversarial networks using representative features. In *Proceedings of the International Conference on Machine Learning*, 2018.
[5] S. Barratt and R. Sharma. A note on the inception score. In *Proceedings of the International Conference on Machine Learning*, 2018.
[6] T. Che, Y. Li, A. P. Jacob, Y. Bengio, and W. Li. Mode regularized generative adversarial networks. *arXiv preprint arXiv:1612.02136*, 2016.
[7] T. Che, Y. Li, A. P. Jacob, Y. Bengio, and W. Li. Mode regularized generative adversarial networks. *CoRR*, 2016.
[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.
[9] C. Doersch. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016.
[10] J. Donahue, P. Krähenbühl, and T. Darrell. Adversarial feature learning. *CoRR*, abs/1605.09782, 2016.
[11] V. Dumoulin, M. I. D. Belghazi, B. Poole, A. Lamb, M. Arjovsky, O. Mastropietro, and A. Courville. Adversarially learned inference. In *Proceedings of the International Conference on Learning Representations*, 2017.
[12] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.
[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.
[14] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, G. Klambauer, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a nash equilibrium. In *Advances in Neural Information Processing Systems*, 2017.
[15] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
[16] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.
[17] A. B. L. Larsen, S. K. Sønderby, and O. Winther. Autoencoding beyond pixels using a learned similarity metric. In *Proceedings of the International Conference on Machine Learning*, pages 1558–1566, 2015.
[18] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. volume 86, pages 2278–2324, 1998.
[19] C. Li, H. Liu, C. Chen, Y. Pu, R. Chen, R. Henao, and L. Carin. Alice: Towards understanding adversarial learning for joint distribution matching. *Neural Information Processing Systems*, 2017.
[20] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision*, 2015.
[21] L. Mescheder, S. Nowozin, and A. Geiger. Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks. In *International Conference on Machine Learning*, 2017.
[22] G. Perarnau, J. van de Weijer, B. Raducanu, and J. M. Álvarez. Invertible conditional gans for image editing. *arXiv preprint arXiv:1611.06355*, 2016.
[23] Y. Pu, W. Wang, R. Henao, L. Chen, Z. Gan, C. Li, and L. Carin. Adversarial symmetric variational autoencoder. In *Neural Information Processing Systems*. 2017.
[24] M. Rosca, B. Lakshminarayanan, D. Warde-Farley, and S. Mohamed. Variational approaches for auto-encoding generative adversarial networks. *arXiv preprint arXiv:1706.04987*, 2017.
[25] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, pages 2234–2242, 2016.
[26] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.
[27] D. Ulyanov, A. Vedaldi, and V. S. Lempitsky. Adversarial generator-encoder networks. *CoRR*, abs/1704.02304, 2017.
[28] D. Ulyanov, A. Vedaldi, and V. S. Lempitsky. It takes (only) two: Adversarial generator-encoder networks. In *AAAI*, 2018.

	CIFAR10 Reconstructions	CIFAR10 Generations	CelebA Reconstructions	CelebA Generations
Approach				
Original				
VAE				
IDVAE				
IDVAE(R)				
VAE-GAN				
α-GAN				

TABLE II: Qualitative experiments comparing different generative models.