IB-GAN: DISENTANGLED REPRESENTATION LEARNING WITH INFORMATION BOTTLENECK GAN

Anonymous authors
Paper under double-blind review

ABSTRACT

We present a novel architecture of GAN for a disentangled representation learning. The new model architecture is inspired by Information Bottleneck (IB) theory thereby named IB-GAN. IB-GAN objective is similar to that of InfoGAN but has a crucial difference; a capacity regularization for mutual information is adopted, thanks to which the generator of IB-GAN can harness a latent representation in disentangled and interpretable manner. To facilitate the optimization of IB-GAN in practice, a new variational upper-bound is derived. With experiments on CelebA, 3DChairs, and dSprites datasets, we demonstrate that the visual quality of samples generated by IB-GAN is often better than those by β-VAEs. Moreover, IB-GAN achieves much higher disentanglement metrics score than β-VAEs or InfoGAN on the dSprites dataset.

1 INTRODUCTION

Learning a *good representation* for data is one of the essential topics in machine learning community. Although any strict definition for it may not exist, general consent about the useful properties of good representation has been discussed throughout many studies (Bengio et al., 2013; Lake et al., 2017). A disentanglement, one of those useful properties of representation, is often described as a statistical independence or factorization; each independent factor is expected to be semantically well aligned with the human intuition on the data generative factors (*e.g.* a chair-type from azimuth on Chairs dataset (Aubry et al., 2014), or age from azimuth on CelebA dataset (Liu et al., 2015)). Learning such disentangled representation, distilling each important attribute of data into a single independent direction, is hard to be done but highly valuable in a variety of other downstream tasks (**?**Higgins et al., 2017b; 2018).

Many models have been proposed for disentangled representation learning (Hinton et al., 2011; Kingma et al., 2014; Reed et al., 2014; Mathieu et al., 2016; Kulkarni et al., 2015; Narayanaswamy et al., 2017; Denton et al., 2017). Despite their impressive results, they either require knowledge of ground-truth generative factors or weak-supervision (*e.g.* domain knowledge or partial labels). In contrast, among many unsupervised approaches (Desjardins et al., 2012; Kingma & Welling, 2013; Rezende et al., 2014; Springenberg, 2015; Dumoulin et al., 2017), yet the two most successful approaches for the independent factor learning are β-VAE (Higgins et al., 2017a) and InfoGAN (Chen et al., 2016).

Higgins et al. (2017a) demonstrate that encouraging the KL-divergence term of Variational autoencoder (VAE) objective (Kingma & Welling, 2013; Rezende et al., 2014) by multiplying a constant $\beta > 1$ induces a high-quality disentanglement of latent factors. As a follow-up research, Burgess et al. (2018) provide a theoretical justification of the principle of β-VAE in terms of the recently popularized Information Bottleneck theory (Tishby & Zaslavsky, 2015; Saxe et al., 2018).

Chen et al. (2016) propose another fully unsupervised approach based on Generative Adversarial Network (GAN) (Goodfellow et al., 2014)). He achieves the goal by enforcing the generator to learn disentangled representation through increasing the mutual information (MI) between the synthetic samples and the latent representation. However, although InfoGAN can learn to disentangle representations for relatively simple datasets (*e.g.* MNIST, 3D Chairs), it struggles to do so on more complicated datasets such as CelebA. Moreover, the representation disentangling performance of InfoGAN is, reportedly known, not as good as β-VAE and its variants with respect to the disentanglement quality metrics proposed in (Higgins et al., 2017a; Kim & Mnih, 2018; Chen et al., 2018).

Stimulated by the success of β-VAE and the IB theory (Burgess et al., 2018) in disentangled representations learning, we hypothesize that the weakness of InfoGAN may originate from that it can only maximize the MI but lacks any constraining mechanism compare to the β-VAE. In other words, InfoGAN misses the term in its objective that acts as β from the perspective of the IB theory.

We present a novel unsupervised model named IB-GAN (*Information Bottleneck GAN*) for learning disentangled representation. We propose a new architecture of GANs from IB theory so that the training objective involves an information capacity constraint that InfoGAN lacks but β-VAE has. We also derive a new variational approximation algorithm to optimize IB-GAN objective in practice. Thanks to the information regularizer, the generator can use the latent representation in a manner that is both more interpretable and disentangled than InfoGAN.

The contributions of IB-GAN over previous disentanglement models are summarized as follows:

1. IB-GAN is a novel GAN-based algorithm for fully unsupervised learning of disentangled representation. To the best of our knowledge, there is no other unsupervised GAN-based model for this sake except the InfoGAN and its variants.
2. Our work is the first attempt to utilize both IB theory and the GAN-based deep generative model. IB-GAN can be seen as an extension to the InfoGAN, supplementing an information constraining regularizer that InfoGAN misses.
3. The visual quality of synthetic samples by IB-GAN is often better than those by β-VAE in our experiments on 3D Chairs (Aubry et al., 2014) and CelebA (Liu et al., 2015) datasets. IB-GAN achieves the state-of-the-art disentanglement scores measured by the metrics of (Higgins et al., 2017a; Eastwood & Williams, 2018) on dSprites dataset (Matthey et al., 2017).

2 PRELIMINARIES

We remind some backgrounds: IB principle in section 2.1 and the relationship between β-VAE and IB theory (Burgess et al., 2018) in section 2.2. Lastly, InfoGAN (Chen et al., 2016) is briefly reviewed in section 2.3.

2.1 INFORMATION BOTTLENECK PRINCIPLE

Let the input variable X and the target variable Y distributed according to some joint data distribution $p(x, y)$. The goal of the IB (Tishby et al., 1999; Alemi et al., 2017; 2018) is to obtain a compressive representation Z from the input variable X, while maintaining the predictive information about the target variable Y as much as possible. The objective for the IB is

$$\max_{q_\phi(z|x)} \mathcal{L}_{\text{IB}} = I(Z, Y) - \beta I(Z, X) \tag{1}$$

where $I(\cdot, \cdot)$ denotes MI and $\beta \geq 0$ is a Lagrange multiplier. The goal is to obtain the optimal representation encoder $q_\phi(z|x)$ that balances the trade-off between the maximization and minimization of both MI terms. Hence, the IB objective in Eq.(1) provides a natural means for *good representation* by enforcing the representation Z to ignore irrelevant information from the input and simultaneously to be predictive about the target, which can act as a minimal sufficient statistic of X for predicting Y (Tishby et al., 1999; Alemi et al., 2017; 2018). In the IB literature, Y typically stands for a classification target label. However, the formulation can be connected to auto-encoding as well (Alemi et al., 2017; 2018; Burgess et al., 2018).

A growing body of studies (Alemi et al., 2017; 2018; Achille & Soatto, 2018; 2017; Burgess et al., 2018; Vera et al., 2017; Rabinowitz et al., 2018) supports that the learned representation from the IB objective contains highly distilled information. For instance, the learned representation is more generalizable and robust to adversarial attack (Alemi et al., 2017), disentangled (Burgess et al., 2018) and invariant to nuance factors (Achille & Soatto, 2017). Moreover, the IB framework prevents weight overfitting (Alemi et al., 2018; Achille & Soatto, 2018; Vera et al., 2017), and can be used to visualize high dimensional embedding in a low dimensional latent space (Rabinowitz et al., 2018).

2.2 β-VAE FROM THE IB THEORY

The key idea of β-VAE (Higgins et al., 2017a) is to multiply a constant $\beta \geq 1$ to the KL-divergence term of the original VAE's objective (Kingma & Welling, 2013; Rezende et al., 2014):

$$\max_{p_\theta, q_\phi} \mathcal{L}_{\beta\text{-VAE}} = \mathbb{E}_{p(x)}[\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - \beta \mathrm{KL}(q_\phi(z|x)||p(z))], \tag{2}$$

where the encoder $q_\phi(z|x)$ is generally known as the variational approximation to the intractable $p(z|x)$, $p(z)$ is a prior for the latent representation and $p_\theta(x|z)$ is the decoder in the VAE context.

Recently, a notable connection between β-VAE and the IB theory has been discovered in (Alemi et al., 2017; 2018). Eq.(2) can be derived from the variational lower-bound approximation to Eq.(1). To clarify this connection, let us see the variational upper and lower bound of the MI:

$$\mathbb{E}_{p(x)}[\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] + H(x) \;\leq\; I_q(Z, X) \;\leq\; \mathbb{E}_{p(x)}[\mathrm{KL}(q_\phi(z|x)||p(z))]. \tag{3}$$

The MI in Eq.(3) subscribed with q: $I_q(Z, X) = E_{q_\phi(z|x)p(x)}[q_\phi(z|x)p(x)/q_\phi(z)p(x)]$, is called as the *representational* MI[1]. Given that computing marginal $q_\phi(z)$ is intractable, we can use any prior $p(z)$ to substitute for $q_\phi(z)$, forming the variational upper-bound in Eq.(3)[2]. Likewise, we can use any decoder model $p_\theta(x|z)$ to approximate $q_\phi(x|z) = q_\phi(z|x)p(x)/q_\phi(z)$ of the MI, forming the variational lower-bound in Eq.(3).

If the target variable Y in Eq.(1) is replaced with X, the task is to reconstruct (auto-encode) data from the representation Z. The variational lower-bound of Eq.(1) can be obtained by leveraging the upper and lower bound of MI in Eq.(3), which corresponds to Eq.(2)[3]. This derivation of β-VAE is more general than the original derivation of VAE using the ELBO (Evidence Lower-Bound) since β naturally arises from the IB objective (Alemi et al., 2017; 2018).

Therefore, the disentanglement-promoting behavior of the β-VAE can be explained in terms of IB theory (Burgess et al., 2018). Constraining the MI (or minimizing KL-divergence in practice) forces the encoder to learn representation containing only strongly relevant information to the data reconstruction, while ignoring other unnecessary (or less-necessary) features. The encoder becomes reluctant to use more channels (or dimensions) of the latent vector to lower the MI constraining cost. Hence, the most distinctive and principle features of data are grouped and aligned along with each independent dimension of the representation space.

2.3 INFOGAN: INFORMATION MAXIMIZING GAN

Generative Adversarial Networks (GAN) (Goodfellow et al., 2014) establish a min-max adversarial game between two neural networks, a generator G and a discriminator D. The discriminator D aims to distinguish well between real sample $x \sim p(x)$ and synthetic sample created by the $G(z)$ with a random noise $z \sim p(z)$, while the generator G is trained to produce a realistic sample that is indistinguishable from the true sample. The adversarial game is formulated as follow:

$$\min_G \max_D \mathcal{L}_{\text{GAN}}(D, G) = \mathbb{E}_{p(x)}[\log(D(x))] + \mathbb{E}_{p(z)}[\log(1 - D(G(z))]. \tag{4}$$

Under an optimal discriminator D^*, Eq.(4) theoretically involves with the Jensen-Shannon divergence between the synthetic and the true sample distribution: $JS(G(z)||p(x))$. However, Eq.(4) does not have any specific guidance on how G utilizes a mapping from z to x. That is, the variation of z in any independent dimension often yields entangled effects on a generated sample x.

On the other hand, InfoGAN (Chen et al., 2016) is capable of learning disentangled representation. InfoGAN introduces an additional latent code c and encourages it to describe the semantic features of the data. To do so, the training objective of InfoGAN accommodates a mutual information maximization term between the latent code c and the generated sample $x = G(z, c)$:

$$\max_G \min_D \mathcal{L}_{\text{InfoGAN}}(D, G) = -\mathcal{L}_{\text{GAN}}(D, G) + \lambda I(c, G(z, c)), \tag{5}$$

where $I(\cdot, \cdot)$ denote MI and λ is a control coefficient. To optimize Eq.(5), the variational lower bound of MI is also exploited similar to that of the IM algorithm (Barber & Agakov, 2003).

[1] We distinguish it from the *generative* in the next section.

[2] The variational inference relies on the positivity of the KL divergence: $\mathbb{E}_{p(\cdot)}[\log p(\cdot)] \geq \mathbb{E}_{p(\cdot)}[\log q(\cdot)]$ for any variational (or approximative) distribution $q(\cdot)$ (Jordan et al., 1999; Wainwright & Jordan, 2008).

[3] A constant data entropy term $H(X) = -\mathbb{E}_{p(x)}[\log p(x)]$ is ignored for brevity.

3 APPROACH

We introduce IB-GAN as a novel disentangled representation learning approach in section 3.1, and propose a practical variational approximation for IB-GAN model in section 3.2. Finally, we discuss some distinctive characteristics of IB-GAN in-depth in section 3.3.

3.1 IB-GAN (INFORMATION BOTTLENECK GAN)

Although InfoGAN (Chen et al., 2016) is a fully unsupervised GAN-based approach for learning disentangled representation, its disentanglement performance is, constantly reported, lower than β-VAE and its variants (Higgins et al., 2017a; Kim & Mnih, 2018; Chen et al., 2018). From the perspective of the IB theory, the weakness of InfoGAN in independent factor learning may originate from the absence of information constraint or any compression mechanism for the representation.

Hence, our motivation is straightforward; we adopt the IB principle to the objective of InfoGAN, presenting Information Bottleneck GAN (IB-GAN). IB-GAN not only maximizes the MI term as the original InfoGAN does, but also constrains the maximization of MI simultaneously as

$$\max_{G} \min_{D} \mathcal{L}_{\text{IB-GAN}}(D,G) = -\mathcal{L}_{\text{GAN}}(D,G) + \lambda I^L(z,G(z)) - \beta I^U(z,G(z)), \quad \text{s.t. } I^L(z,G(z)) \leq I_g(z,G(z)) \leq I^U(z,G(z)), \tag{6}$$

where $I^L(\cdot,\cdot)$ and $I^U(\cdot,\cdot)$ denote the lower and upper bound of *generative* MI[4] respectively. λ and β are the control coefficients (*i.e.* $\lambda \geq \beta$ to satisfy the maximization of MI as such InfoGAN). One important change in Eq.(6) compared to the InfoGAN objective is regularizing the upper bound of MI with β[5]. Hence, this new formulation not only maximizes the shared information between the generator and the representation, but also allows to adjust the maximum amount of information shared by them, analogously to that of β-VAE or IB theory.

3.2 OPTIMIZATION OF IB-GAN

For the optimization of IB-GAN, we here define the tractable variational lower and upper bound of the MI in Eq.(6) using the similar form in (Chen et al., 2016; Agakov & Barber, 2006). For notational consistency, we use $p_\theta(x|z)$ to denote the generator $G(z)$. Then, the variational lower-bound $I^L(z,G(z))$ of the generative MI in Eq.(6) becomes

$$I^L(z,G(z)) = \mathbb{E}_{p_\theta(x|z)p(z)}[\log \frac{q_\phi(z|x)}{p(z)}] \leq I_g(z,G(z)) = \mathbb{E}_{p_\theta(x|z)p(z)}[\log \frac{p_\theta(x|z)p(z)}{p_\theta(x)p(z)}]. \tag{7}$$

Since the generator marginal $p_\theta(x)$ is difficult to calculate, a reconstructor model $q_\phi(z|x)$ is introduced to approximate the quantity $p_\theta(z|x) = p_\theta(x|z)p(z)/p_\theta(x)$ in Eq.(7). The lower-bound holds thanks to positivity of KL-divergence. Intuitively, by improving the reconstruction of an input code z from a generated sample $x = G(z)$, we can maximize the lower-bound of MI between the generator and the code z.

In contrast to the lower-bound, obtaining a practical variational upper-bound of the generative MI is not trivial. If we follow the same approach in (Alemi et al., 2017; 2018), the upper-bound $I^U(z,G(z))$ of the generative MI becomes

$$I_g(z,G(z)) = \mathbb{E}_{p_\theta(x|z)p(z)}[\log \frac{p_\theta(x|z)\cancel{p(z)}}{p_\theta(x)\cancel{p(z)}}] \leq I^U(z,G(z)) = \mathbb{E}_{p_\theta(x|z)p(z)} \log[\frac{p_\theta(x|z)}{d(x)}], \tag{8}$$

where $d(x)$ is a variational approximation to the generator marginal $p_\theta(x) = \sum_z p(x|z)p(z)$. However, one critical problem of this approach is, in practice, it is difficult to choose or correctly identify the proper model (or *distortion* prior in the IB context) for $d(x)$.

[4]The generative MI is described as $I_g(Z,X) = E_{p_\theta(x|z)p(z)}[p_\theta(x|z)p(z)/p_\theta(x)p(z)]$. This initial formulation of the MI is also exhibited in InfoGAN and IM algorithm (Chen et al., 2016; Barber & Agakov, 2003).

[5]We also omit the distinction between z and c for brevity. Hence, z in Eq.(6) is similar to that of c in Eq.(5).

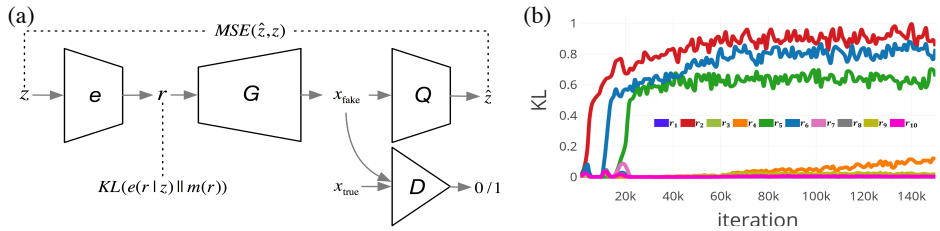

Figure 1: (a) An illustration of IB-GAN architecture. (b) Variations of individual KL-divergence $KL(e(r_i|z)||m(r_i))$ over iterations on dSprites dataset. The sum of these values is the $I^U(z, R(z))$.

Algorithm 1 IB-GAN training algorithm

Input: batch size B, hyperparameters λ, β, and learning rates η_ϕ, η_θ, η_ψ, η_w

while not converged **do**

Sample $\{z^1, \dots, z^B\} \sim p(z)$

Sample $\{x^1, \dots, x^B\} \sim p(x)$

Sample $\{r^1, \dots, r^B\} \sim e_\psi(r|z^i)$ for $i \in \{1 \dots B\}$

Sample $\{x_g^1, \dots, x_g^B\} \sim p_\theta(x|r^i)$ for $i \in \{1 \dots B\}$

$g_\phi \leftarrow \nabla_\phi \frac{1}{B} \sum_i \beta KL(e_\psi(r|z^i)||m(r))$

$g_w \leftarrow -\nabla_w \frac{1}{B} \sum_i \log \sigma(D_w(x_g^i)) + \log(1 - \sigma(D_w(x^i))$

$g_\theta \leftarrow \nabla_\theta \frac{1}{B} \sum_i D_w(x_g^i) - \frac{1}{B} \sum_i \lambda \log q_\phi(z^i|x_g^i)$

$g_\psi \leftarrow \nabla_\psi \frac{1}{B} \sum_i D_w(x_g^i) - \frac{1}{B} \sum_i \lambda \log q_\phi(z^i|x_g^i) + \frac{1}{B} \sum_i \beta KL(e_\psi(r|z^i)||m(r))$

$\phi \leftarrow \phi - \eta_\phi g_\phi;\ w \leftarrow w - \eta_w g_w;\ \theta \leftarrow \theta - \eta_\theta g_\theta;\ \psi \leftarrow \psi - \eta_\psi g_\psi$

end while

In theory, we can choose any distortion model for $d(x)$ (*e.g.* Gaussian), yet any improper choice of $d(x)$ may severely downgrade the quality of synthesized samples from the generator $p_\theta(x|z)$ since the upper-bound $I^U(z, G(z))$ in Eq.(8) is eventually identical to the $KL(p_\theta(x|z)||d(x))$. Moreover, although we express $G(z)$ as $p_\theta(x|z)$ for notional convenience, the probabilistic modeling of generator G will lose the merit of GAN: the likelihood-free (or implicit) modeling assumption.

For this reason, we develop another formulation of the variational upper-bound on the MI term, based on the studies of deep-learning architecture and IB theory (Tishby & Zaslavsky, 2015; Achille & Soatto, 2017; 2018). We define an additional stochastic model $e_\psi(r|z)$ that takes a noise input vector z and produces an intermediate stochastic representation r. In other words, we let $x = G(r(z))$ instead of $x = G(z)$, then we can express the generator as $p_\theta(x|z) = \sum_r p_\theta(x|r)e_\psi(r|z)$. Consequently, a practical upper-bound $I^U(z, R(z))$ of the generative MI can be obtained as:

$$I_g(z, G(R(z))) \leq\ I(z, R(z))\ \big(\leq I(z, z) = H(z)\big) \tag{9}$$

$$I(z, R(z)) = \mathbb{E}_{e_\psi(r|z)p(z)}[\log \frac{e_\psi(r|z)\cancel{p(z)}}{e_\psi(r)\cancel{p(z)}}] \leq I^U(z, R(z)) = \mathbb{E}_{e_\psi(r|z)p(z)} \log[\frac{e_\psi(r|z)}{m(r)}] \tag{10}$$

The first inequality in Eq.(9) holds thanks to the Markov property (Tishby & Zaslavsky, 2015): if any generative process follows $Z \to R \to X$, then $I(Z, X) \leq I(Z, R)$. The inequality in Eq.(10) holds from the positivity of KL divergence. Thus, any distortion prior $m(r)$ can be utilized for substituting the marginal $e_\psi(r)$ without affecting the generated samples directly; therefore, this can bypass the difficulty of choosing distortion prior $d(x)$ in Eq.(8).

Finally, from the variational lower-bound of the MI in Eq.(7) and the newly introduced upper-bound in Eq.(10), the lower-bound of IB-GAN objective in Eq.(6) can be written as:

$$\max_{G, q_\phi, e_\psi} \min_{D} \tilde{\mathcal{L}}_{\text{IB-GAN}}(D, G, q_\phi, e_\psi) = -\mathcal{L}_{\text{GAN}}(D, G) \tag{11}$$
$$+ \lambda \mathbb{E}_{p(z)}[\mathbb{E}_{p_\theta(x|r)e_\psi(r|z)}[\log q_\phi(z|x)] - \beta \text{KL}(e_\psi(r|z)||m(r))]$$

In other words, the intermediate representation r and the $KL(e_\psi(r|z)||m(r))$ with β in Eq.(11) are leveraged to constrain the amount of shared information between the generator $G(z)$ and input z. Eq.(11) is optimized by alternatively maximizing the generator $G = p_\theta(x|r)$, the representation encoder $e_\psi(r|z)$, the variational reconstructor $q_\phi(z|x)$ and the discriminator D. The IB-GAN architecture is presented in Figure 1(a), and overall training procedure is described in Algorithm 1.

3.3 DISCUSSION

The role of the intermediate stochastic model. The IB-GAN involves the stochastic encoder $e_\psi(r|z)$ before the generator to constrain the MI between the generator and input noise z. By doing so, upper-bounding the MI becomes possible without directly affecting the quality of generated samples. Interestingly, if we use an identity mapping function as the encoder model, the natural upper-bound of the MI reduces to the entropy of the input noise distribution $p(z)$ as in Eq.(9). That is, it can be a natural upper-bound for the original InfoGAN and controlled by changing the distribution or the dimension of z. In contrast, IB-GAN can softly adjust the upper-bound by the compression mechanism. In this sense, the encoder can be seen as a hierarchical trainable prior.

Figure 1(b) shows the variations of $\mathrm{KL}(e(r_i|z)||m(r_i))$ for 10-dimensional r (*i.e.* $i = 1, \ldots, 10$) over training iterations on dSprites dataset (Matthey et al., 2017). The sum of these values is the upper-bound of MI. We observe that all factors of variations are capped by different values. It explains that, during training, the encoder $e_\psi(r|z)$ is slowly adapted to the dataset as the upper-bound of MI increases smoothly. This behavior is reported as a key element of the disentangled representation learning in β-VAE (Burgess et al., 2018). Therefore, the generator in IB-GAN learns to parsimoniously utilize each dimension of representation r, resulting in a good disentangled representation learning.

Reconstruction of input noise z. The resulting architecture of IB-GAN is partly analogous to that of β-VAE since both are derived from the IB theory. However, the beta-VAE often generates blurry output images, mainly because it applies the MSE loss to x and uses $\beta > 1$ (Kim & Mnih, 2018; Chen et al., 2018). On the other hand, IB-GAN does not suffer from this shortcoming, because it reconstructs noise z instead of x (*i.e.* applying the MSE loss to z), and uses large $\lambda > \beta$ to maximize the MI. Moreover, it does not rely on any probabilistic modeling assumption of the decoder unlike VAEs and can inherit all merits of InfoGANs (*e.g.* producing images of good quality by an implicit decoder and MI maximization).

Other aspects of IB-GAN. Since the representation encoder $e_\psi(r|z)$ is stochastic, reparametrization trick (Kingma & Welling, 2013) is needed to backpropagate a gradient signal for training the encoder model. The representation r can be embedded along with an extra discrete code $c \sim p(c)$ before getting into the generator (*i.e.* $G(r, c)$), and accordingly the reconstructor network becomes $q(r, c|x)$ to predict the discrete code c as well. In this way, it is straightforward to introduce a discrete representation into IB-GAN, which is not an easy task in β-VAE based models.

If β is too large such that the KL-divergence term is almost zero, then there will be no difference between the samples from the representation encoder $e_\psi(r|z)$ and the distortion prior $m(r)$. Then, both representation r and generated data x contain no information about z at all, resulting in that the signal from the reconstructor is meaningless to the generator. In this case, the IB-GAN reduces to a vanilla GAN with an input $r \sim p(r)$. The similar situation can also happen when $m(r)$ and $p(z)$ are the same distribution (*i.e.* vanilla GAN with $z \sim p(z)$). Therefore, either the dimension or the type of distribution of representation r and noise z have to be different. In our experiments, we use a smaller r dimension than that of noise z.

The disentangled representation in IB-GAN is learned via the representation encoder $e_\psi(r|z)$. To obtain the representation r back from the real data x, we first sampel z using the learned reconstructor $q_\phi(z|x)$, and input it to the representation encoder $e_\psi(r|z)$. For training stability, we also exploit the techniques introduced in PacGAN (Lin et al., 2017); the input layer size of the discriminator is doubled such that the discriminator can see at least two samples at a time to compare them for partly preventing mode collapse.

Related work. Many extensions of β-VAE(Higgins et al., 2017a) have been proposed. Burgess et al. (2018) modify β-VAE's objective such that the KL term is minimized to a specific target constant C instead of scaling the term using β. Kim & Mnih (2018) and Chen et al. (2018) demonstrate using the ELBO surgery (Hoffman & Johnson, 2016; Makhzani & Frey, 2017) that minimizing the KL-divergence enforces factorization of the marginal encoder, and thus promotes the independence of learned representation. However, a high value of β can decrease the MI term too much, and thus often leads to worse reconstruction fidelity compared to the standard VAE. Hence, they introduce a total correlation (Ver Steeg, 2017) based regularization to overcome the reconstruction and disentanglement trade-off. These approaches could be complementary to IB-GAN, since the objective of IB-GAN also involves with the KL term. This exploration could be an interesting future work.

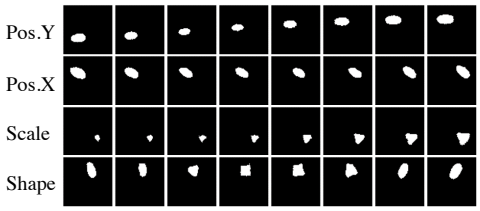

Figure 2: Latent traversals captured by IB-GAN on the dSprites dataset. From top to bottom, each row corresponds to the factors of position Y, position X, scale and shapes.

Table 1: Comparison between methods with the disentanglement metrics in (Kim & Mnih, 2018; Eastwood & Williams, 2018). Our model's scores are obtained from 32 random seeds, with a peak score of (0.826, 0.74). The baseline scores except InfoGAN are referred to (Esmaeili et al., 2018).

Models	**Kim**	**Eastwood**
VAE($\beta = 1.0$)	0.63 ± 0.06	0.30 ± 0.10
β-VAE($\beta = 4.0$)	0.63 ± 0.10	0.41 ± 0.11
β-TCVAE($\beta = 4.0$)	0.62 ± 0.07	0.29 ± 0.10
HFVAE($\beta = 4.0, \gamma = 3.0$)	0.63 ± 0.08	0.39 ± 0.16
InfoGAN($\lambda = 0.05$)	0.59 ± 0.70	0.41 ± 0.05
IB-GAN($\beta = 1, \lambda = 160$)	$\mathbf{0.72 \pm 0.06}$	$\mathbf{0.66 \pm 0.03}$

4 EXPERIMENTS

We experiment our model on various datasets. For quantitative evaluation, we compare methods using the disentanglement metrics proposed in (Kim & Mnih, 2018; Eastwood & Williams, 2018) on dSprites dataset (Matthey et al., 2017) (section 4.1). For qualitative evaluation, we visualize latent traversal results on CelebA (Liu et al., 2015) and 3D Chairs (Aubry et al., 2014) (section 4.2).

We use DCGAN (Radford et al., 2016) with batch normalization (Ioffe & Szegedy, 2015) as our base model for the generator and the discriminator. We let the reconstructor share the same front-end feature with the discriminator for efficient use of parameters as in the InfoGAN (Chen et al., 2016). Also, the MLP-based representation encoder is used before the generator. We train the model using Adam (Kingma & Ba, 2014) optimizer with $\beta_1 = 0.9, \beta_1 = 0.999$. The minibatch size is 64 in all experiments. Lastly, we constrain true and synthetic images to be normalized as $[-1, 1]$. Almost identical architectural configurations for the generator, discriminator, reconstructor, and representation encoder is used in all experiments except that the numbers of parameters is changed depending on the datasets. We defer more details on the models and experimental settings to Appendix.

4.1 QUANTITATIVE RESULTS ON DSPRITES

Although it is not easy to evaluate the disentangled representation, some quantitative metrics (Higgins et al., 2017a; Kim & Mnih, 2018; Chen et al., 2018; Eastwood & Williams, 2018) have been proposed based on the synthetic datasets providing ground-truth generative factors such as dSprites (Matthey et al., 2017) or teapots (Eastwood & Williams, 2018). We verified our approach with the two different metrics (Kim & Mnih, 2018; Eastwood & Williams, 2018) on the dSprites dataset since this setting is tested with many other baselines in (Esmaeili et al., 2018) including standard VAE (Kingma & Welling, 2013; Rezende et al., 2014), β-VAE (Higgins et al., 2017a), TC-VAE (Chen et al., 2018) and HFVAE (Esmaeili et al., 2018).

In experiments, we adopt the instance noises technique (Snderby et al., 2017) since the dSprites images are too simple for the generator of GAN to learn. That is, the intensity distribution of synthetic images is unnaturally narrow (*i.e.* $[0, 255]$), making the overlapping probability with generated images very low, where the generator may barely learn from the discriminator. Hence, by adding instance noises $\epsilon \sim N(0, 255)$ to both true and generated inputs, we can significantly improve the training stability of GAN models. This may be the reason for the inconsistency between the previous experiments (Kim & Mnih, 2018; Chen et al., 2018) on InfoGAN and our results. For the technical detail, we anneal the instance noises linearly from 1 to 0 during training for InfoGAN and IB-GAN.

Table 1 shows the quantitative results in terms of the two disentanglement metric scores (Kim & Mnih, 2018; Eastwood & Williams, 2018). IB-GAN outperforms all other baselines, supporting the effectiveness of proposed method. Interestingly, in our experiments, InfoGAN attains comparable scores to those of other VAE-based models. On the Eastwood's randomforest metric, InfoGAN slightly outperforms other baselines as well, which is consistent with the result of Eastwood & Williams (2018). Lastly, we present the visual inspection of the latent traversal (Higgins et al., 2017a) with the learned IB-GAN model in Figure 2. The IB-GAN successfully learns 4 out of 5 ground truth factors from dSprites dataset, including positions of Y and X, scales and shapes, which aligns with the caps on KL in Figure 1.(b).

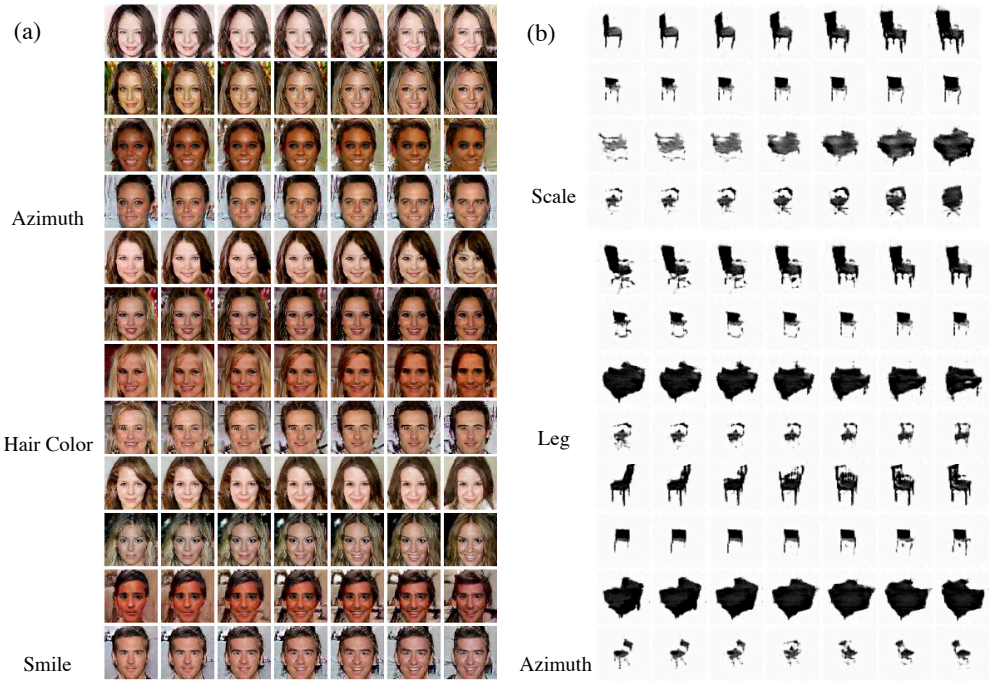

Figure 3: Qualitative results of IB-GAN on latent traversals. (a) IB-GAN learns Azimuth, hair color and smile attributes on the CelebA dataset. (b) it captures factors of scale, whells, and azimuth on the 3D Chairs dataset.

4.2 Qualitative Results on CelebA and 3D Chairs

Following (Chen et al., 2016; Higgins et al., 2017a; Chen et al., 2018; Kim & Mnih, 2018), we evaluate the qualitative results of IB-GAN by inspecting latent traversals. As shown in Figure 3(a), the IB-GAN discovers various human attributes such as Azimuth, hair color and smiling face expression. In addition, generated images of the IB-GAN are sharp and realistic than the result of β-VAE and its variants (Higgins et al., 2017a; Kim & Mnih, 2018; Chen et al., 2018). We also show our qualitative results on 3D Chairs dataset in Figure 3(b). IB-GAN successfully disentangles scales, leg types and azimuth shapes of chairs. These attributes are hardly captured in the original InfoGAN (Chen et al., 2016; Higgins et al., 2017a; Kim & Mnih, 2018; Chen et al., 2018), demonstrating the effectiveness of our approach.

5 Conclusion

The proposed IB-GAN is a novel unsupervised GAN-based model for learning disentangled representation. We made a crucial modification on the InfoGAN's objective inspired by the IB theory and β-VAE; specifically, we developed an information capacity constraining term between the generator and the latent representation. We also derived a new variational approximation technique for optimizing IB-GAN. Our experimental results showed that IB-GAN achieved the state-of-the-art performance on disentangled representation learning. The qualitatively generated samples of IB-GAN often had better quality than those of β-VAE on CelebA and 3D Chairs. IB-GAN attained higher quantitative scores than β-VAE and InfoGAN with disentanglement metrics on dSprites dataset.

There are many possible directions for future work. First, our model can be naturally extended to adapt a discrete latent representation, as discussed in section 3.3. Second, many extensions of β-VAE have been actively proposed such as (Burgess et al., 2018; Kim & Mnih, 2018; Chen et al., 2018; Esmaeili et al., 2018), most of which are complementary for the IB-GAN objective. Further exploration toward this direction could be another interesting next topic.

REFERENCES

Alessandro Achille and Stefano Soatto. Emergence of invariance and disentangling in deep representations. *ArXiv preprint arXiv:1706.01350*, 2017.

Alessandro Achille and Stefano Soatto. Information dropout: Learning optimal representations through noisy computation. *PAMI*, 2018.

Felix V. Agakov and David Barber. Kernelized infomax clustering. In *NIPS*, 2006.

Alexander Alemi, Ben Poole, Ian Fischer, Joshua Dillon, Rif A Saurous, and Kevin Murphy. Fixing a broken elbo. In *ICLR*, 2018.

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep Variational Information Bottleneck. In *ICLR*, 2017.

Mathieu Aubry, Daniel Maturana, Alexei Efros, Bryan Russell, and Josef Sivic. Seeing 3d chairs: exemplar part-based 2d-3d alignment using a large dataset of cad models. In *CVPR*, 2014.

David Barber and Felix Agakov. The IM algorithm: a variational approach to information maximization. In *NIPS*, 2003.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *PAMI*, 2013.

Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-VAE. *ArXiv preprint arXiv:1804.03599*, 2018.

Tian Qi Chen, Xuechen Li, Roger Grosse, and David Duvenaud. Isolating Sources of Disentanglement in Variational Autoencoders. In *NIPS*, 2018.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets. In *NIPS*, 2016.

Emily L Denton et al. Unsupervised learning of disentangled representations from video. In *NIPS*, 2017.

Guillaume Desjardins, Aaron C Courville, and Yoshua Bengio. Disentangling Factors of Variation via Generative Entangling. *ArXiv preprint arXiv:1210.5474*, 2012.

Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Alex Lamb, Martin Arjovsky, Olivier Mastropietro, and Aaron C Courville. Adversarially learned inference. In *ICLR*, 2017.

Cian Eastwood and Christopher K. I. Williams. A framework for the quantitative evaluation of disentangled representations. In *ICLR*, 2018.

Babak Esmaeili, Hao Wu, Sarthak Jain, N Siddharth, Brooks Paige, and Jan-Willem van de Meent. Structured Disentangled Representations. *ArXiv preprint arXiv:1804.02086*, 2018.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C Courville, and Yoshua Bengio. Generative Adversarial Nets. In *NIPS*, 2014.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2017a.

Irina Higgins, Arka Pal, Andrei Rusu, Loic Matthey, Christopher Burgess, Alexander Pritzel, Matthew Botvinick, Charles Blundell, and Alexander Lerchner. DARLA: Improving zero-shot transfer in reinforcement learning. In *ICML*, 2017b.

Irina Higgins, Nicolas Sonnerat, Loic Matthey, Arka Pal, Christopher P Burgess, Matko Bonjak, Murray Shanahan, Matthew Botvinick, Demis Hassabis, and Alexander Lerchner. SCAN: Learning hierarchical compositional visual concepts. In *ICLR*, 2018.

Geoffrey E Hinton, Alex Krizhevsky, and Sida D Wang. Transforming auto-encoders. In *ICANN*, 2011.

Matthew D. Hoffman and Matthew J. Johnson. ELBO surgery: yet another way to carve up the variational evidence lower bound. In *NIPS*, 2016.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015.

Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. An introduction to variational methods for graphical models. *ML*, 1999.

Hyunjik Kim and Andriy Mnih. Disentangling by Factorising. In *ICML*, 2018.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2014.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2013.

Diederik P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In *NIPS*, 2014.

Tejas D Kulkarni, William F Whitney, Pushmeet Kohli, and Josh Tenenbaum. Deep convolutional inverse graphics network. In *NIPS*, 2015.

Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *BBS*, 2017.

Zinan Lin, Ashish Khetan, Giulia Fanti, and Sewoong Oh. PacGAN: The power of two samples in generative adversarial networks. *ArXiv preprint arXiv:1712.04086*, 2017.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015.

Alireza Makhzani and Brendan J Frey. PixelGAN Autoencoders. In *NIPS*. 2017.

Michael F Mathieu, Junbo Jake Zhao, Junbo Zhao, Aditya Ramesh, Pablo Sprechmann, and Yann LeCun. Disentangling factors of variation in deep representation using adversarial training. In *NIPS*, 2016.

Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dSprites: Disentanglement testing Sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017.

Siddharth Narayanaswamy, T. Brooks Paige, Jan-Willem van de Meent, Alban Desmaison, Noah Goodman, Pushmeet Kohli, Frank Wood, and Philip Torr. Learning Disentangled Representations with Semi-Supervised Deep Generative Models. In *NIPS*. 2017.

Neil Rabinowitz, Frank Perbet, Francis Song, Chiyuan Zhang, S. M. Ali Eslami, and Matthew Botvinick. Machine Theory of Mind. In *ICML*, 2018.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. In *ICLR*, 2016.

Scott Reed, Kihyuk Sohn, Yuting Zhang, and Honglak Lee. Learning to disentangle factors of variation with manifold interaction. In *ICML*, 2014.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. In *ICML*, 2014.

Andrew Michael Saxe, Yamini Bansal, Joel Dapello, Madhu Advani, Artemy Kolchinsky, Brendan Daniel Tracey, and David Daniel Cox. On the information bottleneck theory of deep learning. In *ICLR*, 2018.

Jost Tobias Springenberg. Unsupervised and semi-supervised learning with categorical generative adversarial networks. In *ICLR*, 2015.

C. Snderby, J. Caballero, L. Theis, W. Shi, and F. Huszr. Amortised MAP Inference for Image Super-resolution. In *ICLR*, 2017.

Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *ITW*, 2015.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. In *The 37th annual Allerton Conf. on Communication, Control, and Computing*, 1999.

Greg Ver Steeg. Unsupervised Learning via Total Correlation Explanation. In *ICML*, 2017.

Matias Vera, Leonardo Rey Vega, and Pablo Piantanida. Compression-Based Regularization with an Application to Multi-Task Learning. *ICML*, 2017.

Martin J Wainwright and Michael I Jordan. *Graphical Models, Exponential Families, and Variational Inference*. Now Publishers Inc., 2008. ISBN 1601981848, 9781601981844.

A SPECIFICATION OF DATSETS

Table 2: The specification of datasets.

Dataset	Specification
dSprites (Higgins et al., 2017a)	737,280 binary 64×64 images of 2D shapes with 5 ground truth factors. Ground truth factors consist of 3 shapes, 6 scales, 40 orientations, and 32 positions of X and Y.
3D Chairs (Aubry et al., 2014)	86,366 gray-scale 64×64 images of 1,393 chair CAD models with 31 azimuth angles and 2 elevation angles.
CelebA (Liu et al., 2015)	202,599 RGB $64 \times 64 \times 3$ images of celebrity faces consisting of 10,177 identities, 5 landmark locations, and 40 binary attributes. We use the cropped version of the dataset.

B HYPER-PARAMETER SETTINGS

Table 3: The hyperparameter settings for IB-GAN in all experiments.

Dataset	Optimizer	Hyper-Parameters	Regularization
dSprites	Adam(β_1=0.9, β_2=0.999), LR G/E/Q 1e-4, D 1e-5 iterations 3e5	nc=1, ngf=16, ndf=16, z_dim=100, r_dim=10, λ=160, β=1	Instance Noise $\sigma_{instance}$ annealed linearly from 1.0 to 0 for 1e5 iterations.
3D Chairs	Adam(β_1=0.9, β_2=0.999), LR G/E/Q 1e-4, D 1e-5 iterations 1e6	nc=1, ngf=32, ndf=32, z_dim=40, r_dim=10, λ=47, β=1	Instance Noise $\sigma_{instance}$ annealed linearly from 1.0 to 0 for 1e5 iterations. Packing with the degree of 2. (Lin et al., 2017).
CelebA	Adam(β_1=0.9, β_2=0.999), LR G/E/Q 1e-4, D 1e-5 iterations 1e6	nc=3, ngf=64, ndf=64, z_dim=30, r_dim=10, λ=40, β=1	Instance Noise $\sigma_{instance}$ annealed linearly from 1.0 to 0 for 1e5 iterations. Packing with the degree of 2 was exploited (Lin et al., 2017).

C IMPLEMENTATION DETAILS

In this section, we summarize some implementation details of the models in our experiments on dSprites dataset, 3D Chairs, and CelebA datasets. Table 4 shows the base architectures of IB-GAN for the generator, discriminator, and encoder, while Table 5 shows those of InfoGAN. Table 3 also presents the hyperparameter settings that we use for the models in all experiments.

Table 4: The base architecture for IB-GAN. See Table 3 for hyperparameter setting.

Generator(G)	**Discriminator(D) / Encoder(Q)**
Input $z \in \mathbb{R}^{\text{z_dim}}$	Input $x \in \mathbb{R}^{64*64*nc}$
FC ngf*2, BN, ReLU	4x4 conv↓, ndf, lReLU
FC ngf, BN, ReLU	4x4 conv↓, ndf*2, BN, lReLU
FC r_dim*2 $\Rightarrow r_\mu, r_{\log \sigma^2} \in \mathbb{R}^{\text{r_dim}}$	4x4 conv↓, ndf*4, BN, lReLU
Reparametrization Trick $\Rightarrow r \in \mathbb{R}^{\text{r_dim}}$	3x3 conv, ndf*4, BN, lReLU
FC ngf*16, BN, ReLU	3x3 conv, ndf*4, BN, lReLU
FC 8*8*ngf*4, BN, ReLU,	8x8 conv, ndf*16, BN, lReLU
3x3 conv ngf*4, BN, ReLU	FC z_dim $\Rightarrow z_{recon} \in \mathbb{R}^{\text{z_dim}}$
3x3 conv ngf*4, BN, ReLU	FC 1 $\Rightarrow DiscriminatorOutput \in \mathbb{R}^1$
4x4 conv↑ ngf*2, BN, ReLU	
4x4 conv↑ ngf, BN, ReLU	
4x4 conv↑ nc, Tanh $\Rightarrow x_{fake} \in \mathbb{R}^{64*64*nc}$	

Table 5: The base architecture for InfoGAN. All hyperparameters are the same with those of IB-GAN except that z_dim=10.

Generator(G)	**Discriminator(D) / Encoder(Q)**
Input $z \in \mathbb{R}^{\text{z_dim}}$	Input $x \in \mathbb{R}^{64*64*nc}$
FC ngf*16, BN, ReLU	4x4 conv↓, ndf, lReLU
FC 8*8*ngf*4, BN, ReLU	4x4 conv↓, ndf*2, BN, lReLU
3x3 conv ngf*4, BN, ReLU	4x4 conv↓, ndf*4, BN, lReLU
3x3 conv ngf*4, BN, ReLU	3x3 conv, ndf*4, BN, lReLU
4x4 conv↑ ngf*2, BN, ReLU	3x3 conv, ndf*4, BN, lReLU
4x4 conv↑ ngf, BN, ReLU	8x8 conv, ndf*16, BN, lReLU
4x4 conv↑ nc, Tanh $\Rightarrow x_{fake} \in \mathbb{R}^{64*64*nc}$	FC z_dim $\Rightarrow z_{recon} \in \mathbb{R}^{\text{z_dim}}$
	FC 1 $\Rightarrow DiscriminatorOutput \in \mathbb{R}^1$