IB-GAN: Disentangled Representation Learning with Information Bottleneck Generative Adversarial Networks

Insu Jeon, [1] **Wonkwang Lee,** [2] **Myeongjang Pyeon,** [1] **Gunhee Kim** [1]

[1] Dept. of Computer Science and Engineering, Seoul National University, Republic of Korea (South)
[2] School of Computing, Korea Advanced Institute of Science and Technology, Republic of Korea (South)
insuj3on@gmail.com, wonkwang.lee@kaist.ac.kr, mjpyeon@vision.snu.ac.kr, gunhee@snu.ac.kr

Abstract

We propose a new GAN-based unsupervised model for disentangled representation learning. The new model is discovered in an attempt to utilize the Information Bottleneck (IB) framework to the optimization of GAN, thereby named IB-GAN. The architecture of IB-GAN is partially similar to that of InfoGAN but has a critical difference; an intermediate layer of the generator is leveraged to constrain the mutual information between the input and the generated output. The intermediate stochastic layer can serve as a learnable latent distribution that is trained with the generator jointly in an end-to-end fashion. As a result, the generator of IB-GAN can harness the latent space in a disentangled and interpretable manner. With the experiments on dSprites and Color-dSprites dataset, we demonstrate that IB-GAN achieves competitive disentanglement scores to those of state-of-the-art β-VAEs and outperforms InfoGAN. Moreover, the visual quality and the diversity of samples generated by IB-GAN are often better than those by β-VAEs and Info-GAN in terms of FID score on CelebA and 3D Chairs dataset.

Introduction

Learning a good representation of data is one of the essential topics in machine learning research. Although the *goodness* of learned representation depends on the task, a general consensus on the useful properties of representation has been discussed through many recent studies (Bengio, Courville, and Vincent 2013; Ridgeway 2016; Achille and Soatto 2018a). A *disentanglement*, one of such useful properties of representation, is often described as statistical independence of the data generative factors, which is semantically well aligned with human intuition (*e.g.*, chair types or leg shapes on Chairs dataset (Aubry et al. 2014) and age or gender on CelebA dataset (Liu et al. 2015)). Distilling each important factor of data into a single independent direction of representation is hard to be done but invaluable for many other downstream tasks (Ridgeway 2016; Achille and Soatto 2018a; Higgins et al. 2017b, 2018).

Recently, various models have been proposed for disentangled representation learning (Hinton, Krizhevsky, and Wang 2011; Kingma et al. 2014; Reed et al. 2014; Mathieu et al. 2016; N et al. 2017; Denton and vighnesh Birodkar 2017; Jha et al. 2018). 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 (Dumoulin et al. 2017; Donahue, Krähenbühl, and Darrell 2016; Chen et al. 2016; Higgins et al. 2017a; Burgess et al. 2018; Kim and Mnih 2018; Chen et al. 2018), the two most popular approaches maybe β-VAE (Higgins et al. 2017a) and InfoGAN (Chen et al. 2016).

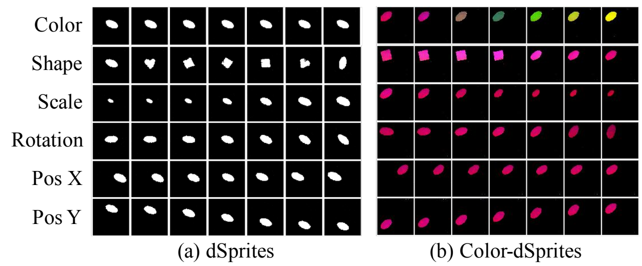


Figure 1: Latent generative factors captured by IB-GAN on (a) dSprites and (b) Color-dSprites dataset. From top to bottom, each row corresponds to the factors of color, shape, scale, rotation, position Y, and position X.

β-VAE (Higgins et al. 2017a) demonstrates that encouraging the KL-divergence term of the Variational Autoencoder (VAE) objective (Kingma and Welling 2014; Rezende, Mohamed, and Wierstra 2014) by multiplying a constant $\beta > 1$ induces strong statistical independence among the factors of latent representation. However, a high β value can strengthen the KL regularization too much, leading to worse reconstruction fidelity than the standard VAE. On the other hand, InfoGAN (Chen et al. 2016) is another fully unsupervised approach based on Generative Adversarial Network (GAN) (Goodfellow et al. 2014). It enforces the generator to learn disentangled representation by increasing the mutual information (MI) between the generated samples and the latent code. Although InfoGAN can learn independent factors well on relatively simple datasets such as MNIST (LeCun, Cortes, and Burges 2010), it struggles to do so on more complicated datasets such as CelebA (Liu et al. 2015) or 3D Chairs (Aubry et al. 2014). Reportedly, the disentangling performance of learned representation by InfoGAN is not as good as that of β-VAE and its variant models (Higgins et al. 2017a; Kim and Mnih 2018; Chen et al. 2018).

Meanwhile, there have been many efforts (Kim and Mnih 2018; Chen et al. 2018; Mathieu et al. 2019) to identify the mechanisms of disentanglement-promoting behavior in β-VAE. Based on the ELBO decomposition (Hoffman and Johnson 2016; Makhzani and Frey 2017), it is revealed that the KL-divergence term in VAE can be factorized to the total correlation term (Watanabe 1960; Hoffman and Johnson 2016), which essentially enforces the factorization of the marginal encoder and thus promotes the independence of learned representations in β-VAE. Besides, some other studies (Burgess et al. 2018; Alemi et al. 2017, 2018) identified that the KL regularization term in β-VAE corresponds to the mutual information (MI) minimization in the variational inference (VI) formulation (Jordan et al. 1999; Wainwright and Jordan 2008) of the Information Bottleneck (IB) theory (Tishby, Pereira, and Bialek 1999; Alemi et al. 2017, 2018).

Based on the aforementioned studies, it is clear that the weakness of GAN-based disentangled representation learning comes from the fact that the model lacks any representation encoder or constraining mechanism for the representation. In the conventional GAN model, a latent representation z is sampled from a fixed latent distribution such as normal distribution, and the generator of GAN maps the whole normal distribution to the target images. Due to this, the latent representation z can be utilized in a highly entangled way; an individual dimension of z to not correspond well to the semantic features of data. Although InfoGAN supports an inverse mapping from data to latent code c, it still does not support disentangled relation as data is inverted to a fixed prior distribution.

In this paper, we present a new GAN-based unsupervised model for disentangled representation learning. Specifically, a new GAN architecture is discovered in an attempt to solve the GAN's objective with IB framework, thereby named IB-GAN (*Information Bottleneck GAN*). The resulting architecture derived from variational inference (VI) formulation of the IB-GAN objective is partially similar to that of InfoGAN but has a critical difference; an intermediate layer of the generator is leveraged to constrain the mutual information between the input and the generated data. The intermediate stochastic layer can serve as a learnable latent distribution that is trained with the generator jointly in an end-to-end fashion. As a result, the generator of IB-GAN can harness the latent space in a disentangled and interpretable manner similar to β-VAE, while inheriting the merit of GANs (*e.g.*, the model-free assumption on generators or decoders, producing good sample quality).

We summarize contributions of this work as follows:

1. IB-GAN is a novel GAN-based model for unsupervised learning of disentangled representation. IB-GAN can be seen as an extension to the InfoGAN, supplementing an information constraining mechanism that InfoGAN lacks in the perspective of IB theory.
2. The resulting IB-GAN architecture derived from the variational inference (VI) formulation of the IB framework supports a trainable latent distribution via intermediate latent encoder between input and the generated data.
3. With the experiments on dSprites (Higgins et al. 2017a) and Color-dSprites dataset (Burgess et al. 2018; Locatello et al. 2019), IB-GAN achieves competitive disentanglement scores to those of state-of-the-art β-VAEs (Burgess et al. 2018; Higgins et al. 2017a; Kim and Mnih 2018; Chen et al. 2018) and outperforms InfoGAN (Chen et al. 2016). The visual quality and diversity of samples generated by IB-GAN are often better than those by β-VAEs and InfoGAN on CelebA(Liu et al. 2015) and 3DChairs (Aubry et al. 2014).

Background

InfoGAN: Information Maximizing GAN

Generative Adversarial Networks (GAN) (Goodfellow et al. 2014) formulate a min-max adversarial game between two neural networks, a generator G and a discriminator D:

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

The discriminator D aims to distinguish well between real samples $x \sim p(x)$ and synthetic samples created by the generator $G(z)$ with random noise $z \sim p(z)$, while the generator G is trained to produce realistic sample that is indistinguishable from true sample. Under an optimal discriminator D^*, Eq.(1) theoretically minimizes the Jensen-Shannon divergence between the synthetic and the true sample distribution: $JS(G(z)||p(x))$. However, Eq.(1) 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.

InfoGAN (Chen et al. 2016) is capable of learning disentangled representations without any supervision. To do so, the objective of InfoGAN accommodates a mutual information maximization term between an additional latent code c and a generated sample $x = G(z, c)$:

$$\begin{aligned}&\min_G \max_D \mathcal{L}_{\text{InfoGAN}}(D, G)\\ &= \mathcal{L}_{\text{GAN}}(D, G) - I(c, G(z, c)), \qquad (2)\end{aligned}$$

where $I(\cdot, \cdot)$ denote MI. Also, c and z are latent code and not interpretable (or in-compressible) noise respectively. To optimize Eq.(2), the variational lower-bound of MI is exploited similarly to the IM algorithm (Barber and Agakov 2004).

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 IB (Tishby, Pereira, and Bialek 1999; Alemi et al. 2017, 2018; Peng et al. 2019) is to obtain a compressive representation Z from the input X, while maintaining the predictive information about the target Y as much as possible. The objective for the IB is

$$\max_{q_\phi(z|x)} \mathcal{L}_{\text{IB}} = I(Z, Y) - \beta I(Z, X), \qquad (3)$$

where $I(\cdot, \cdot)$ denotes MI and $\beta \geq 0$ is a Lagrange multiplier. Therefore, IB aims at obtaining the optimal representation encoder[1] $q_\phi(z|x)$ that simultaneously balances the

[1] ϕ is the parameter of representation encoder model.

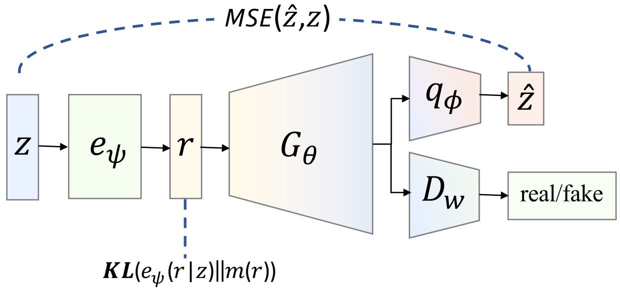


Figure 2: An illustration of IB-GAN. It adopts a representation encoder $e_\psi(r|z)$ and a KL-divergence loss derived from IB theory. Since the encoder $q_\phi(r|z)$ is assumed as Gaussian, it is convenient to define $m(r)$ as Gaussian. The MSE loss is computed by the decoder $\log q_\phi(z|x)$ in Eq.(11).

tradeoff between the maximization and minimization of both MI terms. Accordingly, the learned representation Z can act as a minimal sufficient statistic of X for predicting Y. The IB principle provides an intuitive meaning for the *good representation* from the perspective of information theory. Recent studies (Burgess et al. 2018; Achille and Soatto 2018a; Zhao, Song, and Ermon 2018) show that the variational inference (VI) formulation (Jordan et al. 1999; Wainwright and Jordan 2008) of IB Eq.(3) is equivalent to the objective of β-VAE when the task is self-reconstruction (*e.g.*, $Y = X$).

IB-GAN: Information Bottleneck GAN

The motivation for IB-GAN is straight forwards. We can identify that InfoGAN's objective in Eq.(2) lacks a MI minimization term compare to the IB objective in Eq.(3). Thus, we utilize the MI minimization term to InfoGAN's objective to get the IB-GAN objective as follows:

$$\begin{aligned}
&\min_G \max_D \mathcal{L}_{\text{IB-GAN}}(D,G) \\
&= \mathcal{L}_{\text{GAN}}(D,G) - \left[I^L(z,G(z)) - \beta I^U(z,G(z))\right], \quad (4)\\
&\quad \text{s.t.} \;\; I^L(z,G(z)) \le I_g(z,G(z)) \le I^U(z,G(z)), \quad (5)
\end{aligned}$$

where $I^L(\cdot,\cdot)$ and $I^U(\cdot,\cdot)$ denote the lower and upper-bound of *generative* MI[2] respectively. One important change in Eq.(4) compared to Eq.(2) is adopting the upper-bound of MI with a trade-off coefficient β, analogously to that of β-VAE and the IB objective[3]. More discussion regarding this parameter is presented in the next section, and the effect is explored in the experimental section.

[2]The *generative* mutual information (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 formulation of MI is also exhibited in InfoGAN and IM algorithm (Chen et al. 2016; Barber and Agakov 2004).

[3]The incompressible noise variable z is not necessarily required for modeling InfoGAN (Srivastava et al. 2017; Zhao, Song, and Ermon 2018). So, we can omit the incompressible noise z. Here, z has the same role as the latent code c in InfoGAN.

Algorithm 1 IB-GAN training algorithm

Input: batch size B, hyperparameters β, and the learning rates η_ϕ, η_θ, η_ψ, and η_w of the parameter of reconstructor, generator, encoder, and discriminator model respectively.
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 $i \in \{1\dots B\}$
  Sample $\{\hat{z}^1,\dots,\hat{z}^B\} \sim q_\phi(z|x_g^i)$ for $i \in \{1\dots B\}$
  $g_\phi \leftarrow \nabla_\phi \frac{1}{B}\sum_i (\hat{z}^i - z^i)^2$
  $g_w \leftarrow -\nabla_w \frac{1}{B}\sum_i \log\sigma(D_w(x_g^i)) + \log(1-\sigma(D_w(x^i)))$
  $g_\theta \leftarrow \nabla_\theta \frac{1}{B}\sum_i \log\sigma(D_w(x_g^i)) - (\hat{z}^i - z^i)^2$
  $g_\psi \leftarrow \nabla_\psi \frac{1}{B}\sum_i \log\sigma(D_w(x_g^i)) - (\hat{z}^i - z^i)^2 + \beta KL(e_\psi(r|z^i)||m(r))$
  $\phi \leftarrow \phi - \eta_\phi g_\phi;\; w \leftarrow w - \eta_w g_w;$
  $\theta \leftarrow \theta - \eta_\theta g_\theta;\; \psi \leftarrow \psi - \eta_\psi g_\psi$
end while

Optimization of IB-GAN

For the optimization of Eq.(4), we first define a tractable lower-bound of the *generative* MI in Eq.(5) using the similar derivation exhibited in (Chen et al. 2016; Agakov and Barber 2005; Alemi and Fischer 2018). For the brevity, we use probabilistic model notion (*i.e.*, $p_\theta(x|z) = \mathcal{N}(G_\theta(z), \mathbf{1})$ for the generator. Then, the variational lower-bound $I^L(z,G(z))$ of the *generative* MI in Eq.(5) is given as

$$\begin{aligned}
&I_g(z,G(z)) = \mathbb{E}_{p_\theta(x|z)p(z)}[\log \frac{p_\theta(x|z)p(z)}{p_\theta(x)p(z)}] \\
&\ge I^L(z,G(z)) = \mathbb{E}_{p_\theta(x|z)p(z)}[\log \frac{q_\phi(z|x)}{p(z)}] \quad (6)\\
&= \mathbb{E}_{p_\theta(x|z)p(z)}[\log q_\phi(z|x)] + H(z). \quad (7)
\end{aligned}$$

In Eq.(6), the lower-bound holds thanks to positivity of KL-divergence. A variational reconstructor $q_\phi(z|x)$ is introduced to approximate the quantity $p_\theta(z|x) = p_\theta(x|z)p(z)/p_\theta(x)$. Intuitively, by improving the reconstruction of an input code z from a generated sample $x = G(z)$ using the $q_\phi(z|x)$, we can promote the statistical dependency between the generator $G(z)$ and the input latent code z (Chen et al. 2016; Barber and Agakov 2004).

In contrast to the lower-bound, obtaining a practical variational upper-bound of the *generative* MI in Eq.(5) is not trivial. If we follow the similar approach discussed in previous studies (Alemi et al. 2017, 2018), the variational upper-bound $I^U(z,G(z))$ of the generative MI is derived as

$$\begin{aligned}
&I_g(z,G(z)) = \mathbb{E}_{p_\theta(x|z)p(z)}[\log \frac{p_\theta(x|z)\cancel{p(z)}}{p_\theta(x)\cancel{p(z)}}] \\
&\le I^U(z,G(z)) = \mathbb{E}_{p_\theta(x|z)p(z)} \log[\frac{p_\theta(x|z)}{d(x)}], \quad (8)
\end{aligned}$$

where $d(x)$ approximates the generator marginal $p_\theta(x) = \sum_z p_\theta(x|z)p(z)$. In theory, we can choose any approximation model $d(x)$. However, one important concern here is that it is difficult to correctly identify a proper approximation model for $d(x)$ in practice. Given that the upper-bound $I^U(z, G(z))$ is identical to $KL(p_\theta(x|z)||d(x))$ in Eq.(8), $d(x)$ acts as an image prior. Thus, any improper choice of $d(x)$ may severely downgrade the quality of synthesized samples from the generator $p_\theta(x|z)$. We might mitigate this by introducing another GAN loss for the KL divergence, but the effective prior model $d(x)$ is still missing.

For this reason, we propose another formulation of the variational upper-bound on the *generative* MI, inspired by the recent studies of deep-learning with IB principle (Tishby and Zaslavsky 2015; Achille and Soatto 2018a,b). 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[4] as $p_\theta(x|z) = \sum_r p_\theta(x|r)e_\psi(r|z)$. Subsequently, a new variational upper-bound $I^U(z, R(z))$ can be obtained as

$$\begin{aligned} & I_g(z, G(R(z))) \\ & \leq I(z, R(z)) = \mathbb{E}_{e_\psi(r|z)p(z)}[\log \frac{e_\psi(r|z)\cancel{p(z)}}{e_\psi(r)\cancel{p(z)}}] \quad (9) \\ & \leq I^U(z, R(z)) = \mathbb{E}_{e_\psi(r|z)p(z)} \log[\frac{e_\psi(r|z)}{m(r)}]. \quad (10) \end{aligned}$$

The first inequality in Eq.(9) holds due to the Markov property (Tishby and Zaslavsky 2015): if any generative process follows $Z \to R \to X$, then $I(Z, X) \leq I(Z, R)$. The second inequality in Eq.(10) holds from the positivity of KL divergence. Any model for $m(r)$ can be flexibly used to approximate the representation marginal $e_\psi(r)$ (*e.g.*, Gaussian). This new formulation of the variational upper-bound in Eq.(10) allows us to constrain the *generative* MI without directly affecting the generator in Eq.(8) via the intermediate representation encoder model $e_\psi(r|z)$ and the prior $m(r)$.

Finally, from the lower-bound in Eq.(7) and the upper-bound in Eq.(10), a variational approximation of Eq.(4) can be obtained as

$$\begin{aligned} & \min_{G, q_\phi, e_\psi} \max_D \tilde{\mathcal{L}}_{\text{IB-GAN}}(D, G, q_\phi, e_\psi) \\ & = \mathcal{L}_{\text{GAN}}(D, G) - \big(\mathbb{E}_{p(z)}[\mathbb{E}_{p_\theta(x|r)e_\psi(r|z)}[\log q_\phi(z|x)] \\ & \quad - \beta \text{KL}(e_\psi(r|z)||m(r))]\big). \quad (11) \end{aligned}$$

We define the encoder $e_\psi(r|z)$ as a stochastic model $N(\mu_\psi(z), \sigma_\psi(z))$ and the prior $m(r)$ as $N(\mathbf{0}, \mathbf{1})$, as done in VAEs (Kingma and Welling 2014). The optimization of Eq.(11) can be done by alternatively updating the parameters of the generator $p_\theta(x|r)$, the representation encoder $e_\psi(r|z)$, the variational reconstructor $q_\phi(z|x)$ and the discriminator D using any SGD-based algorithm. A reparameterization trick (Kingma and Welling 2014) is employed to backpropagate gradient signals to the stochastic encoder. The overall architecture of IB-GAN is presented in Figure 2, and the training procedure is described in Algorithm 1.

Related Work and Discussion

Connection to rate-distortion theory. IB theory is a generalization of the Rate-Distortion (RD) theory (Tishby, Pereira, and Bialek 1999), in which the rate $\mathcal{R}$ is the code length per data sample to be transmitted through a noisy channel, and the distortion $\mathcal{D}$ represents the approximation error of reconstructing the input from the source code (Alemi et al. 2018; Shannon 1948). The goal of RD-theory is to minimize $\mathcal{D}$ without exceeding a certain level of rate $\mathcal{R}$, formulated as $\min_{\mathcal{R},\mathcal{D}} \mathcal{D} + \beta\mathcal{R}$, where $\beta \in [0, \infty]$ decides a theoretically achievable optimum in the auto-encoding limit (Alemi et al. 2018).

IB-GAN can be described in terms of the RD-theory. Here, the goal is to deliver an input code z through a noisy channel (*i.e.*, deep neural networks). Both r and x are regarded as the encoding of input z. The distortion in IB-GAN corresponds to the reconstruction error of the input z estimated from the variational encoder $q_\phi(z|x(r))$.

The rate $\mathcal{R}$ of the intermediate representation r is related to $KL(e_\psi(r|z)||m(r))$, which measures the inefficiency (or the excess rate) of the representation encoder $e_\psi(r|z)$ in terms of how much it deviates from the approximating representation prior $m(r)$. Hence, β in Eq.(11) controls the compressing level of the information contained in r for reconstructing input z. It constrains the amount of shared information between the input code z and the output image by the generator $x = G(r(z))$ without directly regularizing the output image itself. In addition, the GAN loss $\mathcal{L}_{\text{GAN}}$ in Eq.(11) can be understood as a rate constraint of the image in the context of RD-theory since the GAN loss approximates $JS(G(z)||p(x))$ (Goodfellow et al. 2014) between the generator and the empirical data distribution $p(x)$.

Comparison between IB-GAN and β-VAE. The resulting architecture of IB-GAN is partly analogous to that of β-VAE since both are derived from the IB theory[5]. β-VAE tends to generate blurry output images due to large β (Kim and Mnih 2018; Chen et al. 2018). Setting β to large value minimizes the excess rate of encoding z in β-VAE, but this also increases the reconstruction error (or the distortion) (Alemi et al. 2018). In contrast, IB-GAN may not directly suffer from this shortcoming of β-VAE. The generator of IB-GAN learns to encode image x by minimizing the rate (*i.e.*, $JS(G(r)||p(x))$) inheriting the merit of InfoGANs (*e.g.*, an implicit decoder model can be trained to produce a good quality of images). One possible drawback of IB-GAN architecture is that it does not directly map the representation encoder to output r from the real image x: $q(r|x)$. To obtain the representation r back from the real data x, we need a two-step procedure: sampling z from the learned reconstructor $q_\phi(z|x)$ and putting it to the representation encoder $e_\psi(r|z)$. However, the latent representation r obtained from this procedure is quite compatible with those of β-VAEs as we will see in the experimental section.

Disentanglement-promoting behavior of IB-GAN. The disentanglement-promoting behavior of β-VAE is encouraged by the KL divergence. Since the prior distribution is often assumed as a fully factored Gaussian distribution, the KL

[4]In this case, we let $p_\theta(x|r) = \mathcal{N}(G_\theta(r), \mathbf{1})$.

[5]IB-GAN's objective is derived from the *generative* MI, while β-VAE's objective is derived from the *representational* MI in (Alemi et al. 2017, 2018).

divergence term in VAE objective can be decomposed into the form containing a total correlation (TC) term (Watanabe 1960; Hoffman and Johnson 2016), which essentially enforces the statistical factorization of the representation (Kim and Mnih 2018; Chen et al. 2018; Burgess et al. 2018). In IB-GAN, a noise z is treated as the input source instead of image x. Therefore, the disentangling mechanism of IB-GAN is slightly different from that of β-VAE.

The disentanglement-promoting behavior of IB-GAN can be described in term of the RD-theory as follow: (1) The efficient encoding scheme for the (intermediate) latent representation r can be learned by minimizing $KL(e_\psi(r|z)||m(r))$ with a factored Gaussian prior $m(r)$, which promotes statistical factorization of the coding r similar to that of VAE. (2) The efficient encoding scheme for x is defined by minimizing the divergence between $G(z)$ and the data distribution $p(x)$ via the discriminator, which promotes the encoding of x to be a realistic image. (3) Maximizing $I^L(z, G(z))$ in IB-GAN indirectly maximizes $I(r, G(r))$ too since $I(z, G(z)) \leq I(r, G(r))$ from the Markov propriety (Tishby and Zaslavsky 2015). That is, maximizing the lower-bound of MI increases the statistical dependency between the coding r and $x = G(r)$, while both encoding r and x need to be efficient in terms of their rates (*e.g.*, the upper-bound of MI and the GAN loss). Therefore, an independent directional change in the representation encoding r tends to be well aligned with a predominant factor of variation in the image x.

Other characterizations of IB-GAN. IB-GAN can softly constrain the *generative* MI by the variational upper-bound derived in Eq.(10). In this regard, the variational encoder of IB-GAN can be seen as a hierarchical trainable prior for the generator. If β in Eq.(11) is zero, the IB-GAN objective reduces to that of InfoGAN. In contrast, if β is too large such that the KL-divergence term is almost zero, then there would 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 the signal from the reconstructor being meaningless to the generator. If we further exclude the lower-bound of MI in Eq.(11), the IB-GAN objective reduces to that of vanilla GAN with an input $r \sim m(r)$.

Variational bounds on generative MI. Maximizing the variational lower-bound of *generative* MI has been employed in IM algorithm (Agakov and Barber 2005) and InfoGAN (Chen et al. 2016). Recently, Alemi and Fischer (Alemi and Fischer 2018) propose the lower-bound of *generative* MI, named GILBO, as a data-independent measure that can quantify the complexity of the learned representations for trained generative models. They discover that the lower-bound is correlated with the image quality metrics of generative models such as INCEPTION (Barratt and Sharma 2018) and FID (Heusel et al. 2017) scores. On the other hand, we propose a new approach of upper-bounding the *generative* MI, based on the causal relationship of deep learning architecture, and show the effectiveness of the upper-bound by measuring the disentanglement scores (Kim and Mnih 2018) on the learned representation.

Experiments

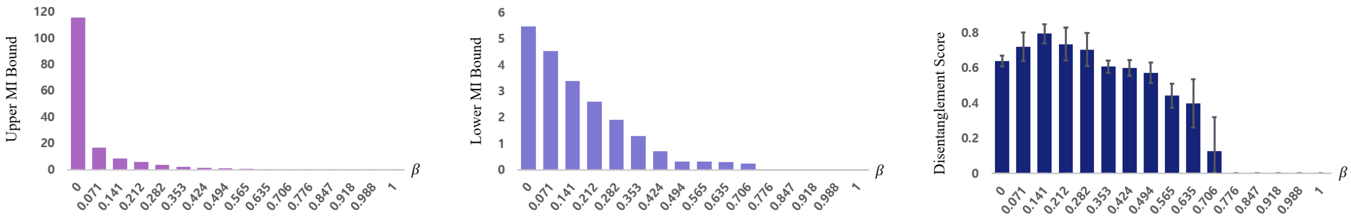
We experiment IB-GAN on various datasets. For quantitative evaluation, we measure the disentanglement metrics proposed in (Kim and Mnih 2018) on dSprites (Higgins et al. 2017a) and Color-dSprites (Burgess et al. 2018; Locatello et al. 2019) dataset. For qualitative evaluation, we visualize latent traversal results of IB-GAN and measure FID scores (Szegedy et al. 2015) on CelebA (Liu et al. 2015) and 3D Chairs (Aubry et al. 2014) dataset.

Architecture. We follow DCGAN (Radford, Metz, and Chintala 2016) with batch normalization (Ioffe and Szegedy 2015) for both generator and discriminator of IB-GAN. We let the reconstructor $q_\phi(z|x)$ share the same front-end features with the discriminator $D(x)$ for the efficient use of parameters as in the conventional InfoGAN (Chen et al. 2016) model. Also, an MLP-based representation encoder $e_\psi(r|z)$ is used before the generator $G(r)$. Optimization is performed with RMSProp (Tieleman and Hinton 2012) with a momentum of 0.9. The batch size is 64 in all experiments. We constrain true and synthetic images to be normalized as $[-1, 1]$. Lastly, we use almost identical architecture for the generator, discriminator, reconstructor, and representation encoder in all of our experiments, except the different sizes of channel parameters depending on the datasets. We defer more details of the IB-GAN architecture to Appendix.

Quantitative Results

Although it is not easy to evaluate the disentanglement of representation, some quantitative metrics (Higgins et al. 2017a; Kim and Mnih 2018; Chen et al. 2018) have been proposed based on the synthetic datasets that provide ground-truth generative factors such as dSprites (Higgins et al. 2017a) or Color-dSprites (Burgess et al. 2018; Locatello et al. 2019). We evaluate our approach with the metric of (Kim and Mnih 2018) on the dSprites and Color-dSprites datasets since many other state-of-the-art models are evaluated in this setting in (Locatello et al. 2019), including standard VAE (Kingma and Welling 2014; Rezende, Mohamed, and Wierstra 2014), β-VAE (Higgins et al. 2017a), TC-VAE (Chen et al. 2018) and FactorVAE (Kim and Mnih 2018).

Disentanglement performance. According to IB (or RD) theory (Alemi et al. 2018), we can set any real values to β. For the quantitative evaluation, we perform hyperparameter search in the range of $\beta \in [0, 1]$. We focus on investigating the effect of $\beta \in [0, 1]$ on the MI and the disentangling promoting behavior. Table 1 compares the disentanglement performance metric of Kim and Mnih (2018) between methods on the dSprites and Color-dSprites (Burgess et al. 2018; Locatello et al. 2019) dataset. The optimal average disentanglement scores 0.80 and 0.79 on the two datasets are obtained at $\beta = 0.141$ and $\beta = 0.071$, respectively. In our experiment, the disentanglement scores of IB-GAN exceed those of GAN (Goodfellow et al. 2014), VAE (Kingma and Welling 2014; Rezende, Mohamed, and Wierstra 2014) and InfoGAN (Chen et al. 2016), and are comparable to those of β-VAE. For the VAE baselines, we follow the model architectures and experimental settings of (Locatello et al. 2019).



(a) $I^U(z, G(z))$ vs β.
(b) $I^L(z, G(z))$ vs β.
(c) Disentanglement score vs β.

Figure 3: Effects of β on the converged upper/lower-bound of MI and disentanglement metric scores (Kim and Mnih 2018).

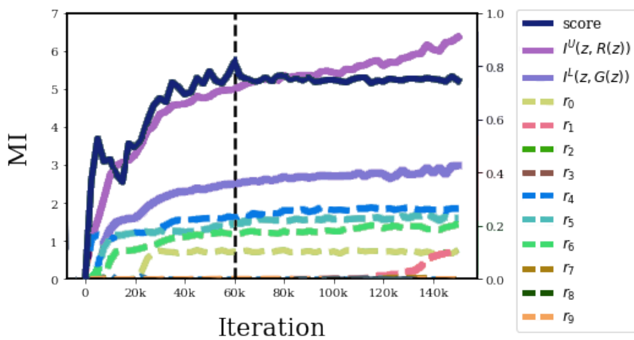


Figure 4: The plot of variational upper-bound and lower-bound of MI with independent $\mathrm{KL}(e(r_i|z)||m(r_i))$ values for all r_i ($i = 1, \dots, 10$) and disentanglement scores (Kim and Mnih 2018) over 150K training iterations. The vertical dashed black line represents the iteration at the highest disentanglement score.

For the GAN baselines, we use the subset of components of IB-GAN: generator and discriminator for the vanilla GAN and additional reconstructor for the InfoGAN.

Traversal examples. Figure 1(a) presents the visual inspection of the latent traversal (Higgins et al. 2017a) with the learned IB-GAN model on dSprites. The IB-GAN successfully learns 5 out of 5 ground-truth factors from the dSprites, including Y and X positions, scales, rotations and shapes, which align well with the KL scores in Figure 4. Figure 1(b) presents that IB-GAN captures 6 out of 6 ground-truth factors additionally including the color factor on Color-dSprites (Burgess et al. 2018; Locatello et al. 2019).

Convergence. Figure 4 shows the variations of $\mathrm{KL}(e(r_i|z)||m(r_i))$ for 10-dimensional r (*i.e.*, $i = 1, \dots, 10$) over training iterations on dSprites when $\beta = 0.212$. The increasing curves indicate that information capture by r_i for x increases as the learned representation is informative to reconstruct the input correctly. Moreover, each r_i increases at different points; it implies that the variational encoder $e_\psi(r|z)$ of IB-GAN slowly adapts to capture the independent factors of variations in dSprites as the upper-bound of MI increases. The similar behavior is reported in β-VAE (Burgess et al. 2018). More results of convergence plots are presented in Appendix.

Models	dSprites	Color-dSprites
GAN	0.40 ± 0.05	0.35 ± 0.04
InfoGAN	0.61 ± 0.03	0.55 ± 0.08
IB-GAN	$\mathbf{0.80 \pm 0.07}$	$\mathbf{0.79 \pm 0.05}$
VAE	0.61 ± 0.04	0.59 ± 0.06
β-VAE	0.69 ± 0.09	0.74 ± 0.06
FactorVAE	$\mathbf{0.81 \pm 0.07}$	$\mathbf{0.82 \pm 0.06}$
β-TCVAE	0.79 ± 0.06	0.80 ± 0.07

Table 1: Comparison of disentanglement metric values (Kim and Mnih 2018). The average scores of IB-GAN is obtained from 10 random seeds.

The effect of β. We inspect the effect of β on the convergence of upper and lower MI bounds and the disentanglement score (Kim and Mnih 2018) on dSprites. We take a median value over the 150K training iterations in each trial, and then average the values over 10 different trials per β in a range of $[0, 1]$. Figure 3(a) and 3(b) illustrate the expected converged value of upper and lower MI bounds over the different β. When $\beta = 0$, the upper MI bound in the IB-GAN objective disappears; hence, the representation encoding r can diverge from the prior distribution $m(r)$ without any restriction, resulting in a high divergence. When a small $\beta > 0$ is set, the MI upper bound constraint affects the optimization procedure. Thus the divergence between r and its prior decreases drastically. After then, the MI upper bound seems to decrease gradually as the β gets larger, consequently the lower MI bound decreases as well. Lastly, Figure 3(c) shows the effect of β on the disentanglement scores. The average disentanglement score varies according to β, supporting that we could control the disentangling-promoting behavior of IB-GAN with the upper-bound of *generative* MI and β. Especially, the optimal disentanglement scores are achieved when β is in a range of $[0.071,\ 0.212]$.

Qualitative Results

Following (Chen et al. 2016; Higgins et al. 2017a; Chen et al. 2018; Kim and Mnih 2018), we evaluate the qualitative results of IB-GAN by inspecting latent traversals. As shown in Figure 5(a), IB-GAN discovers various human recognizable attributes such as azimuth, gender, and skin tone on CelebA dataset. We also present the results of IB-GAN on 3D Chairs in Figure 5(b), where IB-GAN disentangles azimuth, scales, and leg types of chairs. These attributes are

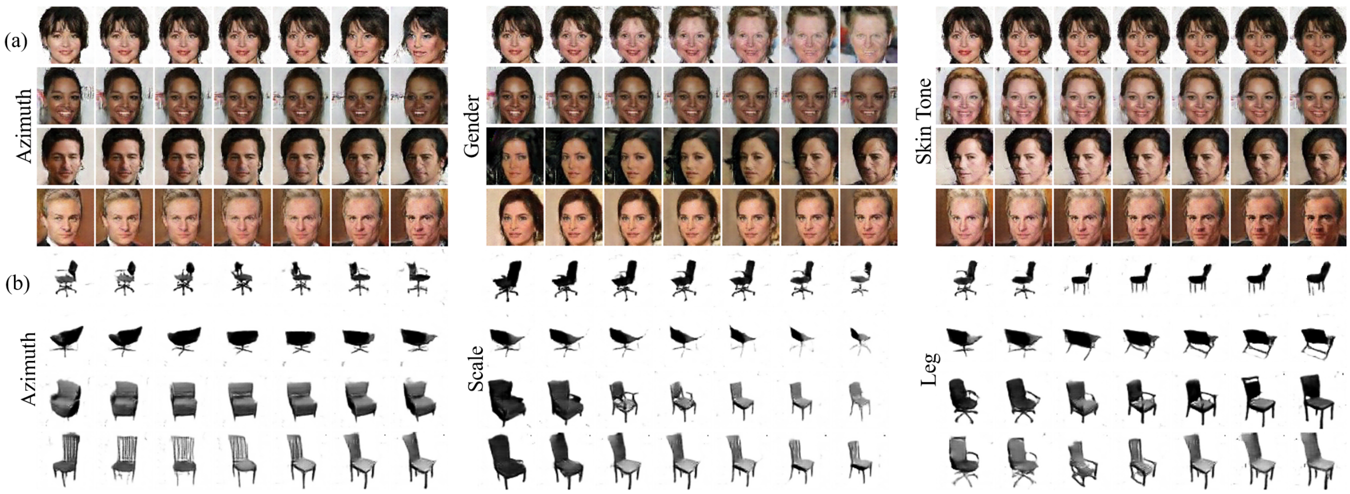


Figure 5: Latent traversals of IB-GAN that captures the factors of (a) azimuth, gender and skin tone attributes on CelebA and (b) scale, leg and azimuth on 3D Chairs. More factors captured by IB-GAN are presented in Appendix.

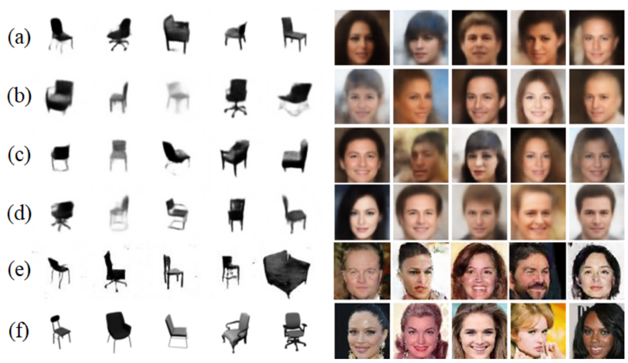


Figure 6: Comparison of random samples on CelebA and 3D Chairs dataset: (a) VAE, (b) β-VAE, (c) FactorVAE, (d) β-TCVAE, and (e) IB-GAN and (f) real images.

hardly captured by the original InfoGAN (Chen et al. 2016; Higgins et al. 2017a; Kim and Mnih 2018; Chen et al. 2018), demonstrating the effectiveness of the proposed model.

Figure 6 illustrates randomly sampled images generated by IB-GAN and the VAE baselines. Figure 6 shows that the images obtained from IB-GAN are often sharper and more realistic than those obtained from β-VAE and its variants (Higgins et al. 2017a; Kim and Mnih 2018; Chen et al. 2018). More qualitative results are presented in Appendix.

FID scores. In Table 2, the FID score (Szegedy et al. 2015) of IB-GAN is significantly lower than those of VAEs and comparable to those of GANs, indicating that the generator of IB-GAN can produce diverse and qualitative image generation, while not only capturing various factors of variations from the dataset. One reason for the generalization performance improvement in IB-GAN is the flexibility of learning the prior distribution. In contrast, other GAN baselines only rely on the pre-specified prior distributions. VAEs tend to degrade the reconstruction of the image due to their strong independent assumption on the representation.

Models	CelebA	3D Chairs
VAE	129.7	56.2
β-VAE	131.0	91.3
FactorVAE	109.7	44.7
β-TCVAE	125.0	57.3
GAN	8.4	27.9
InfoGAN	9.3	25.6
IB-GAN	**7.4**	**25.5**

Table 2: FID scores on CelebA and 3D Chairs dataset. The lower FID score, the better quality, and diversity of samples.

Conclusion

The proposed IB-GAN model is a new unsupervised GAN-based model for disentangled representation learning. Inspired by IB theory, we employ the MI minimization term to InfoGAN's objective to get the IB-GAN objective. The resulting architecture derived from the variational inference (VI) formulation of IB-GAN's objective is partially similar to that of InfoGAN but has a critical difference; an intermediate layer of the generator is leveraged to constrain the mutual information between the input and the generated data. The intermediate stochastic layer can serve as a learnable latent representation distribution that is trained with the generator jointly in an end-to-end fashion. As a result, the generator of IB-GAN can harness the latent space in a disentangled and interpretable manner similar to β-VAE, while inheriting the merit of GANs (*e.g.*, the model-free assumption on generators or decoders, producing good sample quality). Our experimental results demonstrate that IB-GAN shows good performance on disentangled representation learning comparable with β-VAEs and outperforms InfoGANs. Moreover, the qualitative results also exhibit that IB-GAN can be trained to generate diverse and high-quality visual samples while capturing various factors of variations on CelebA and 3D Chairs dataset.

Acknowledgments

This work was supported by Center for Applied Research in Artificial Intelligence(CARAI) grant funded by Defense Acquisition Program Administration(DAPA) and Agency for Defense Development(ADD) (UD190031RD). Gunhee Kim is the corresponding author. We would like to thank Byeongchang Kim and Youngjae Yu for helpful comments.

References

Achille, A.; and Soatto, S. 2018a. Emergence of invariance and disentanglement in deep representations. *JMLR* 19(1): 1947–1980.

Achille, A.; and Soatto, S. 2018b. Information dropout: Learning optimal representations through noisy computation. *PAMI* 40(12): 2897–2905.

Agakov, F. V.; and Barber, D. 2005. Kernelized Infomax Clustering. In *NeurIPS*, 17–24.

Alemi, A. A.; and Fischer, I. 2018. GILBO: one metric to measure them all. In *NeurIPS*, 7037–7046.

Alemi, A. A.; Fischer, I.; Dillon, J.; and Murphy, K. 2017. Deep Variational Information Bottleneck. In *ICLR*.

Alemi, A. A.; Poole, B.; Fischer, I.; Dillon, J.; Saurous, R. A.; and Murphy, K. 2018. Fixing a Broken ELBO. In *ICML*, volume 80, 159–168.

Aubry, M.; Maturana, D.; Efros, A. A.; Russell, B. C.; and Sivic, J. 2014. Seeing 3D Chairs: Exemplar Part-based 2D-3D Alignment using a Large Dataset of CAD Models. In *CVPR*.

Barber, D.; and Agakov, F. 2004. The IM algorithm: a variational approach to information maximization. In *NeurIPS*, volume 16, 201.

Barratt, S.; and Sharma, R. 2018. A Note on the Inception Score. *arXiv preprint arXiv:1801.01973* .

Bengio, Y.; Courville, A.; and Vincent, P. 2013. Representation learning: A review and new perspectives. *PAMI* 35(8): 1798–1828.

Burgess, C. P.; Higgins, I.; Pal, A.; Matthey, L.; Watters, N.; Desjardins, G.; and Lerchner, A. 2018. Understanding disentangling in β-VAE. *arXiv preprint arXiv:1804.03599* .

Chen, R. T. Q.; Li, X.; Grosse, R. B.; and Duvenaud, D. K. 2018. Isolating Sources of Disentanglement in Variational Autoencoders. In *NeurIPS*, volume 31.

Chen, X.; Duan, Y.; Houthooft, R.; Schulman, J.; Sutskever, I.; and Abbeel, P. 2016. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In *NeurIPS*, 2180–2188.

Denton, E. L.; and vighnesh Birodkar. 2017. Unsupervised Learning of Disentangled Representations from Video. In *NeurIPS*, volume 30.

Donahue, J.; Krähenbühl, P.; and Darrell, T. 2016. Adversarial feature learning. In *ICLR*.

Dumoulin, V.; Belghazi, I.; Poole, B.; Mastropietro, O.; Lamb, A.; Arjovsky, M.; and Courville, A. 2017. Adversarially Learned Inference. In *ICLR*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative Adversarial Nets. In *NeurIPS*, volume 27.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. GANs trained by a two time-scale update rule converge to a Nash equilibrium. In *NeurIPS*, volume 30.

Higgins, I.; Matthey, L.; Pal, A.; Burgess, C.; Glorot, X.; Botvinick, M.; Mohamed, S.; and Lerchner, A. 2017a. β-VAE: Learning basic visual concepts with a constrained variational framework. In *ICLR*.

Higgins, I.; Pal, A.; Rusu, A.; Matthey, L.; Burgess, C.; Pritzel, A.; Botvinick, M.; Blundell, C.; and Lerchner, A. 2017b. DARLA: Improving Zero-Shot Transfer in Reinforcement Learning. In *ICML*, volume 70, 1480–1490.

Higgins, I.; Sonnerat, N.; Matthey, L.; Pal, A.; Burgess, C. P.; Bosnjak, M.; Shanahan, M.; Botvinick, M.; Hassabis, D.; and Lerchner, A. 2018. SCAN: Learning Hierarchical Compositional Visual Concepts. In *ICLR*.

Hinton, G. E.; Krizhevsky, A.; and Wang, S. D. 2011. Transforming auto-encoders. In *ICANN*.

Hoffman, M. D.; and Johnson, M. J. 2016. ELBO surgery: yet another way to carve up the variational evidence lower bound. In *NeurIPS*, volume 1, 2.

Ioffe, S.; and Szegedy, C. 2015. Batch Normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, volume 37, 448–456.

Jha, A. H.; Anand, S.; Singh, M.; and Veeravasarapu, V. 2018. Disentangling factors of variation with cycle-consistent variational auto-encoders. In *ECCV*, 805–820.

Jordan, M. I.; Ghahramani, Z.; Jaakkola, T. S.; and Saul, L. K. 1999. An Introduction to Variational Methods for Graphical Models. *ML* 37(2): 183–233.

Kim, H.; and Mnih, A. 2018. Disentangling by factorising. In *ICML*, 2649–2658.

Kingma, D. P.; Mohamed, S.; Jimenez Rezende, D.; and Welling, M. 2014. Semi-Supervised Learning with Deep Generative Models. In *NeurIPS*, volume 27.

Kingma, D. P.; and Welling, M. 2014. Auto-Encoding Variational Bayes. In *ICLR*.

LeCun, Y.; Cortes, C.; and Burges, C. 2010. MNIST handwritten digit database. *ATT Labs* 2. URL http://yann.lecun.com/exdb/mnist/.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep learning face attributes in the wild. In *ICCV*, 3730–3738.

Locatello, F.; Bauer, S.; Lucic, M.; Raetsch, G.; Gelly, S.; Schölkopf, B.; and Bachem, O. 2019. Challenging common assumptions in the unsupervised learning of disentangled representations. In *ICML*, 4114–4124. PMLR.

Makhzani, A.; and Frey, B. J. 2017. PixelGAN Autoencoders. In *NeurIPS*, volume 30.

Mathieu, E.; Rainforth, T.; Siddharth, N.; and Teh, Y. W. 2019. Disentangling disentanglement in variational autoencoders. In *ICML*, 4402–4412.

Mathieu, M. F.; Zhao, J. J.; Zhao, J.; Ramesh, A.; Sprechmann, P.; and LeCun, Y. 2016. Disentangling factors of variation in deep representation using adversarial training. In *NeurIPS*.

N, S.; Paige, B.; van de Meent, J.-W.; Desmaison, A.; Goodman, N.; Kohli, P.; Wood, F.; and Torr, P. 2017. Learning Disentangled Representations with Semi-Supervised Deep Generative Models. In *NeurIPS*, volume 30.

Peng, X. B.; Kanazawa, A.; Toyer, S.; Abbeel, P.; and Levine, S. 2019. Variational Discriminator Bottleneck: Improving Imitation Learning, Inverse RL, and GANs by Constraining Information Flow. In *ICLR*.

Radford, A.; Metz, L.; and Chintala, S. 2016. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. In *ICLR*.

Reed, S.; Sohn, K.; Zhang, Y.; and Lee, H. 2014. Learning to Disentangle Factors of Variation with Manifold Interaction. In *ICML*, volume 32, 1431–1439.

Rezende, D. J.; Mohamed, S.; and Wierstra, D. 2014. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. In *ICML*, volume 32, 1278–1286.

Ridgeway, K. 2016. A Survey of Inductive Biases for Factorial Representation-Learning. *arXiv preprint arXiv:1612.05299* .

Shannon, C. E. 1948. A mathematical theory of communication. *The Bell system technical journal* 27(3): 379–423.

Srivastava, A.; Valkov, L.; Russell, C.; Gutmann, M. U.; and Sutton, C. 2017. VEEGAN: Reducing Mode Collapse in GANs using Implicit Variational Learning. In *NeurIPS*, volume 30.

Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; and Rabinovich, A. 2015. Going Deeper with Convolutions. In *CVPR*, 1–9.

Tieleman, T.; and Hinton, G. 2012. Lecture 6.5—RmsProp: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks for Machine Learning.

Tishby, N.; Pereira, F. C.; and Bialek, W. 1999. The information bottleneck method. In *The 37th annual Allerton Conference on Communication, Control, and Computing*, 368–377.

Tishby, N.; and Zaslavsky, N. 2015. Deep learning and the information bottleneck principle. In *Information Theory Workshop (ITW)*, 1–5.

Wainwright, M. J.; and Jordan, M. I. 2008. *Graphical Models, Exponential Families, and Variational Inference*. Now Publishers Inc.

Watanabe, S. 1960. Information theoretical analysis of multivariate correlation. *IBM Journal of research and development* 4(1): 66–82.

Zhao, S.; Song, J.; and Ermon, S. 2018. The Information Autoencoding Family: A Lagrangian Perspective on Latent Variable Generative Models. In *UAI*.