Supplementary Material for "IB-GAN: Disengangled Representation Learning with Information Bottleneck Generative Adversarial Networks"

Anonymous authors

Paper 9686 under double-blind review

Keywords: Generative Adversarial Network, Information Bottleneck, Unsupervised Representation Learning

A. Related Work

A.1 β-VAE

β-VAE (Higgins et al. 2017) is one of the state-of-the-art models for unsupervised disentangled representation learning. The key idea of β-VAE is to multiply the KL-divergence term of the original VAE's objective (Kingma and Welling 2013; Rezende, Mohamed, and Wierstra 2014) by a constant $\beta \geq 1$:

$$\max_{p_\theta, q_\phi} \mathcal{L}_{\beta\text{-VAE}} = \mathbb{E}_{p(x)}[\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)]] - \beta \mathrm{KL}(q_\phi(z|x)||p(z))], \tag{1}$$

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

Recently, the connection between the β-VAE and the Information Bottleneck (IB) theory has been discovered in (Alemi et al. 2017, 2018a). That is, Eq.(1) is equivalent to the variational formulation of IB objective[1]. Given that computing the marginal of mutual information (MI) in the IB objective is intractable, the variational lower and upper-bound based on the *representational* MI[2] is derived as:

$$I_q(Z,Y) \geq \mathbb{E}_{p(y)}[\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(y|z)] + H(y), \quad I_q(Z,X) \leq \mathbb{E}_{p(x)}[\mathrm{KL}(q_\phi(z|x)||p(z))]. \tag{2}$$

$p(z)$ is used to approximate $q_\phi(z)$, forming the variational upper-bound[3], while $p_\theta(x|z)$ approximates $q_\phi(x|z) = q_\phi(z|x)p(x)/q_\phi(z)$, forming the variational lower-bound of the MI. Since the target in VAE is to reconstruct data X from the representation Z, we can set X instead of the target variable Y in Eq.(2). Consequently, the variational lower-bound of IB objective, obtained from the lower and upper-bound of the MI in Eq.(2), corresponds to the β-VAE's objective in Eq.(1).

A.2 GANs

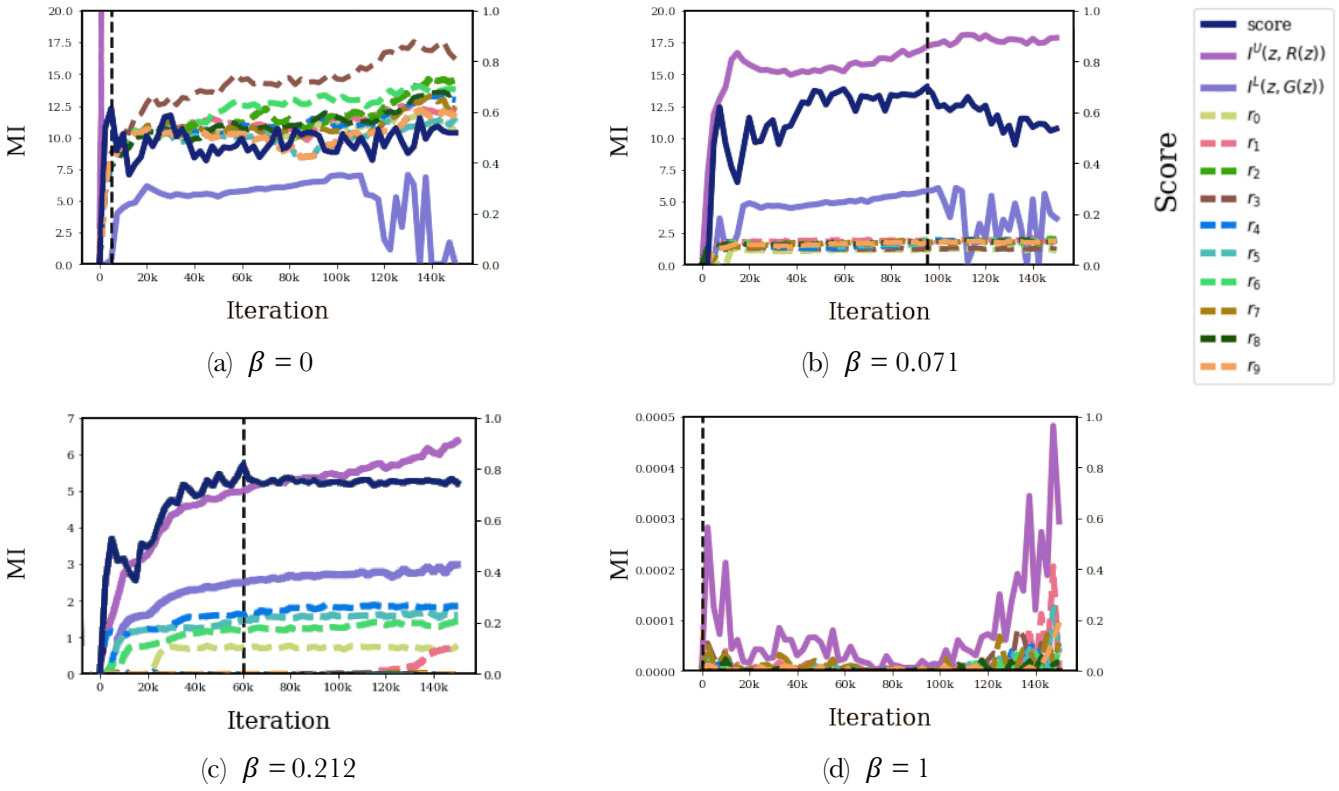
There have been many studies for GAN-based representation learning in a (semi-) supervised way (Kulkarni et al. 2015; Reed et al. 2014; Narayanaswamy et al. 2017; Mathieu et al. 2016) or an unsupervised way (Springenberg 2016; Dumoulin et al. 2017; Donahue, Krähenbühl, and Darrell 2017). InfoGAN is an unsupervised GAN model that is dedicated to the disentangled representation learning. Nevertheless, its disentanglement quality based on the Kim's metric (Kim and Mnih 2018) has been reported less comparable to that of the β-VAE (Higgins et al. 2017; Kim and Mnih 2018; Chen et al. 2018). We have extended InfoGAN to IB-GAN (*Information Bottleneck GAN*) by adding the upper-bound of *generative* MI into the InfoGAN's objective. IB-GAN inherits several advantages of GANs (*e.g.* the model-free assumption on generators, producing good quality of samples, and the potential use of discrete latent variables).

---

[1] The IB objective is the Eq.(3) in the manuscript (i.e. $\mathcal{L}_{\text{IB}} = I(Z,Y) - \beta I(Z,X)$).

[2] The mutual information (MI) based on the encoder $q_\phi(z|x)$ is referred to the *representational* MI in (Alemi et al. 2018b) (i.e. $I_q(Z,X) = E_{q_\phi(z|x)p(x)}[q_\phi(z|x)p(x)/q_\phi(z)p(x)]$). We distinguish it from the *generative* MI based on the generator $p_\theta(x|z)$ described in the manuscript.

[3] The variational inference technique 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 approximating) distribution $q(\cdot)$ (Jordan et al. 1999; Wainwright and Jordan 2008).



(a) $\beta = 0$ (b) $\beta = 0.071$

(c) $\beta = 0.212$ (d) $\beta = 1$

Figure 1: Effects of β on the convergence of variational upper-bound and lower-bound of MI. We also depict individual KL-term $\text{KL}(e(r_i|z)||m(r_i))$ (dashed lines) for each r_i ($i = 1, \cdots, 10$) over 150K training iterations. Note that the sum of all independent KL divergences is the upper-bound of MI (i.e. $I^U(z, R(z)) = \sum_i \text{KL}(e(r_i|z)||m(r_i))$). Each vertical dashed black line indicates the iteration at the highest disentanglement scores (Kim and Mnih 2018).

B. Additional Experiments

B.1 The effects of β in IB-GAN on dSprites dataset

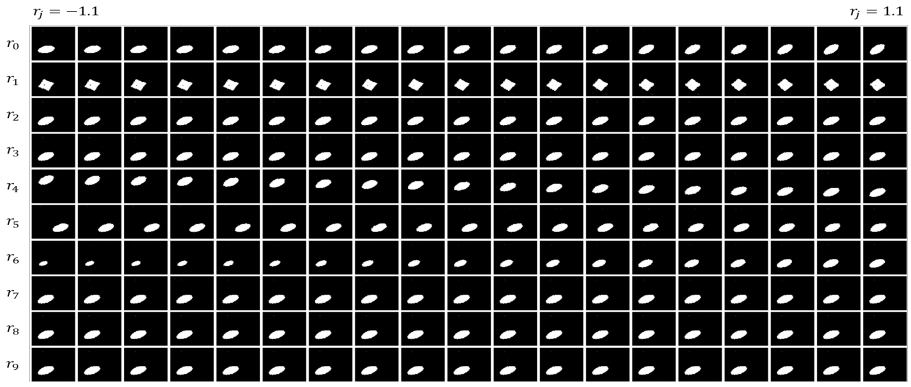
To investigate the effect of β, we illustrate the convergence of the upper and lower MI bounds and the disentanglement scores in Figure 1. The KL divergence for each independent dimension of the representation is also displayed in Figure 1.

When $\beta = 0$ as shown in Figure 1(a), the constraining effect of the upper-bound of MI is disappeared. Hence, it is hard to distinguish the information levels captured on each independent representation. In this case, IB-GAN only has the power of maximizing the lower MI bound, similar to InfoGAN. As a result, the disentanglement score is not relatively high.

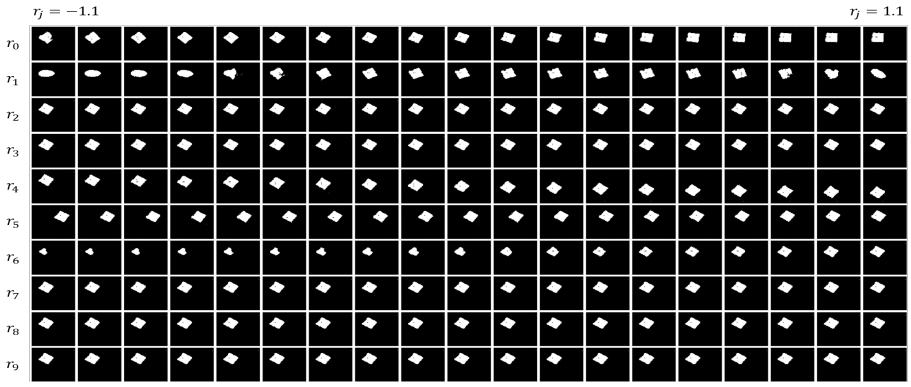
On the other hand, when β is 0.212, as in Figure 1(c), both the lower and upper-bound of MI increases smoothly. The representation encoder $e_\psi(r|z)$ is slowly learned to capture the dataset's distinctive factors. Each independent KL-divergence of the representation is capped by different values. A similar behavior is observed as a key element of the disentangled representation learning in β-VAE (Burgess et al. 2018). Therefore, the generator in IB-GAN with the proper β value can learn to parsimoniously utilize each dimension of representation r, resulting in a good disentangled representation learning.

When $\beta = 1$ as in Figure 1(d), the upper-bound of MI drops down to almost zero, and so does the lower-bound of MI due to the constraining effect of the upper-bound. In this case, the behavior of IB-GAN is similar to the standard GAN, which yields entangled representation.

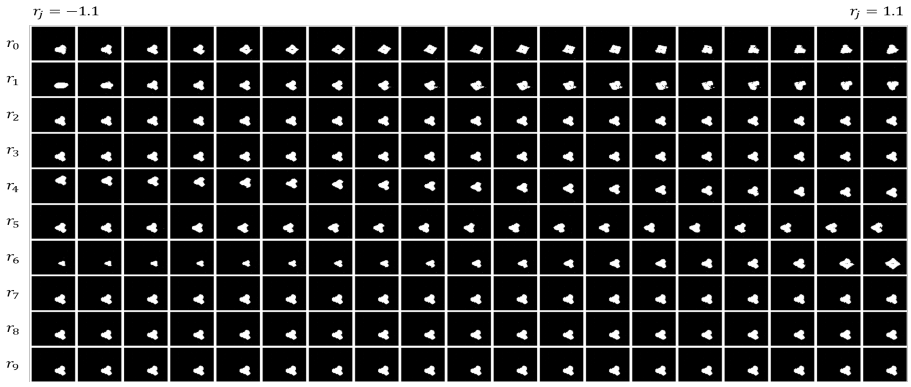
B.2 Latent traversal samples on dSprites dataset



(a) Ellipse



(b) Square

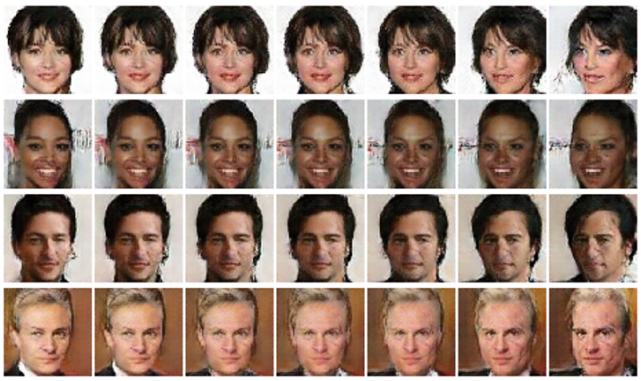


(c) Heart

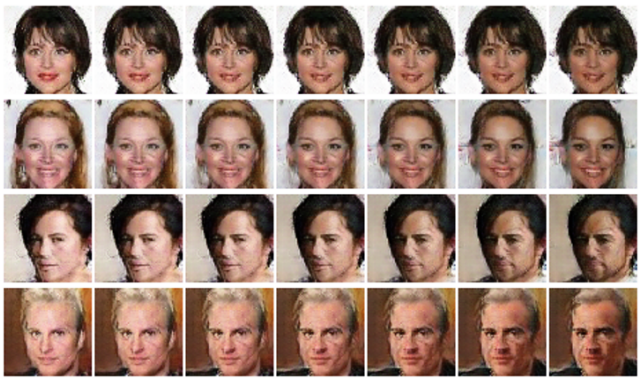
Figure 2: Some examples of latent traversals of three different base shapes (ellipse, square, and heart) on dSprites with the best parameter setting ($\beta = 0.212$). IB-GAN successfully captures the five factors of variations: rotations (r_0), shapes (r_1), positions of Y (r_4) and X (r_5) and scales (r_6). The generator does not reflect the changes in r_2, r_3, r_7, r_8 and r_9 since they are identical to factored zero-mean Gaussian prior $m(r_i)$ and convey no information about z. These results align with Figure.1(c); the KL-divergence values of these dimensions are nearly zero.

B.3 Latent traversal samples on CelebA dataset

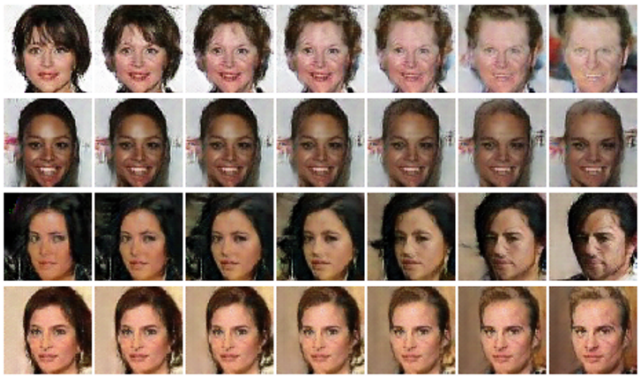
We illustrate more qualitative results of IB-GAN trained on CelebA datasets. As shown in Figure 3, IB-GAN discovers various attributes of human faces: (a) azimuth, (b) skin tone, (c) gender, (d) smile, (e) hair length and (f) hair color. All features in Figure 3 are captured by the model trained with the parameter setting of $\beta = 0.325, \gamma = 2$. Note that these attributes are hardly captured in the original InfoGAN (Chen et al. 2016; Higgins et al. 2017; Kim and Mnih 2018; Chen et al. 2018), demonstrating the effectiveness of disentangled representation learning by IB-GAN.



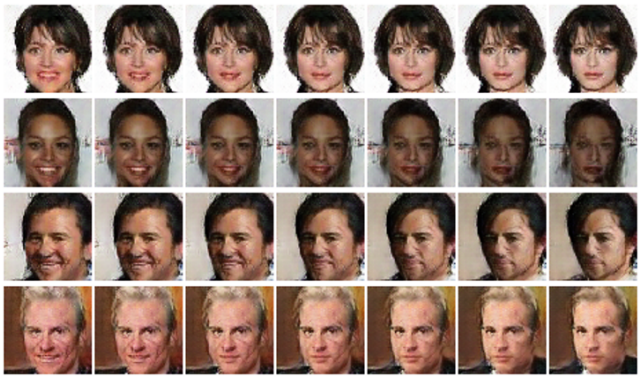
(a) Azimuth



(b) Skin Tone

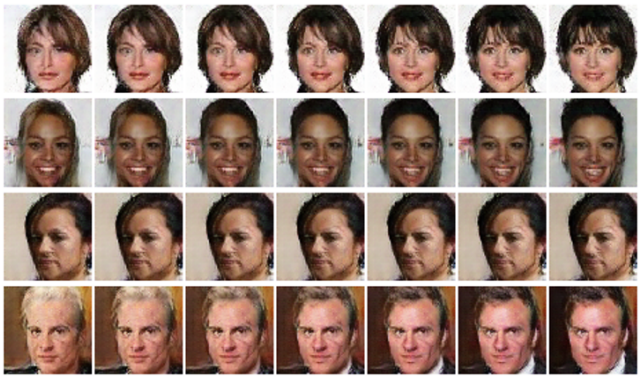


(c) Gender



(d) Smile

(e) Hair Length



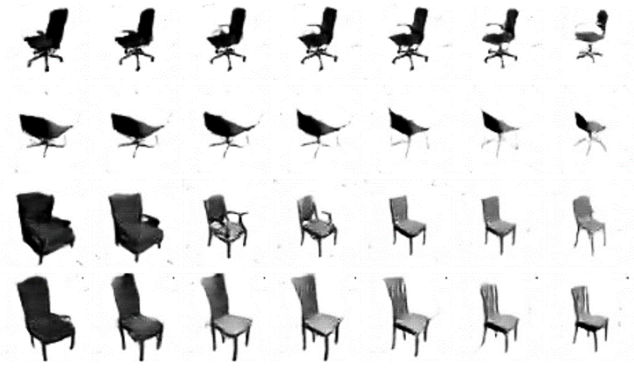
(f) Hair Color

Figure 3: Latent traversals of attributes captured by six different r vectors on CelebA dataset with the parameter setting of $\beta = 0.325, \gamma = 2$.

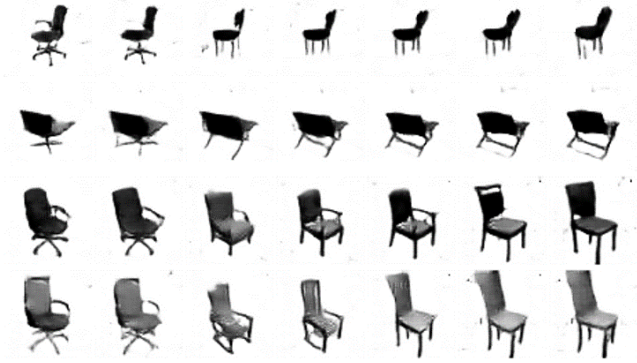
B.4 Latent traversal samples on 3D Chair dataset

For 3D Chairs dataset, Figure 4 shows that IB-GAN can capture following factors of chairs: (a) azimuth, (b) scale, (c) leg, (d) back length and (e) width. We obtain the results with the parameter setting of $\beta = 0.35$ and $\gamma = 1.2$.

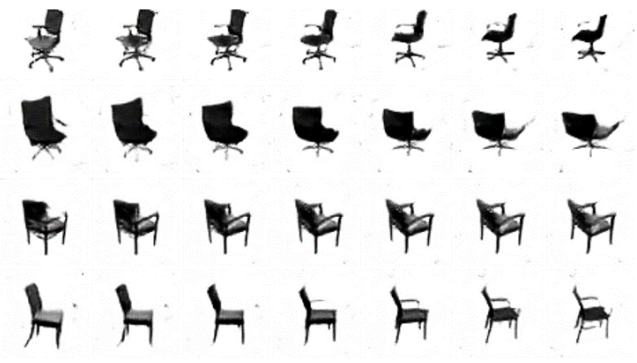
(a) Azimuth



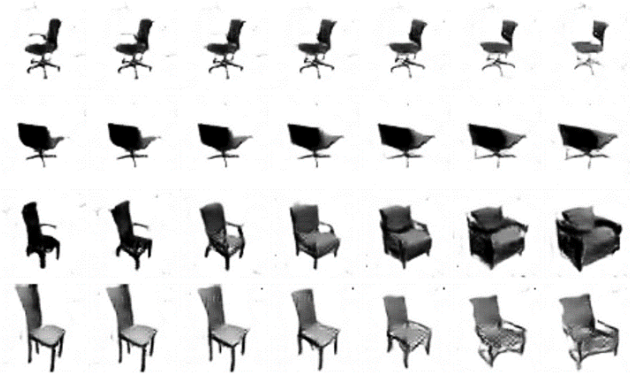
(b) Scale



(c) Leg



(d) Back Length



(e) Width

Figure 4: Latent traversals of attributes captured by five different r vectors on 3D Chairs dataset with the parameter setting of $\beta = 0.35$, $\gamma = 1.2$.

B.5 Random samples generated by the models on CelebA and 3D Chairs dataset

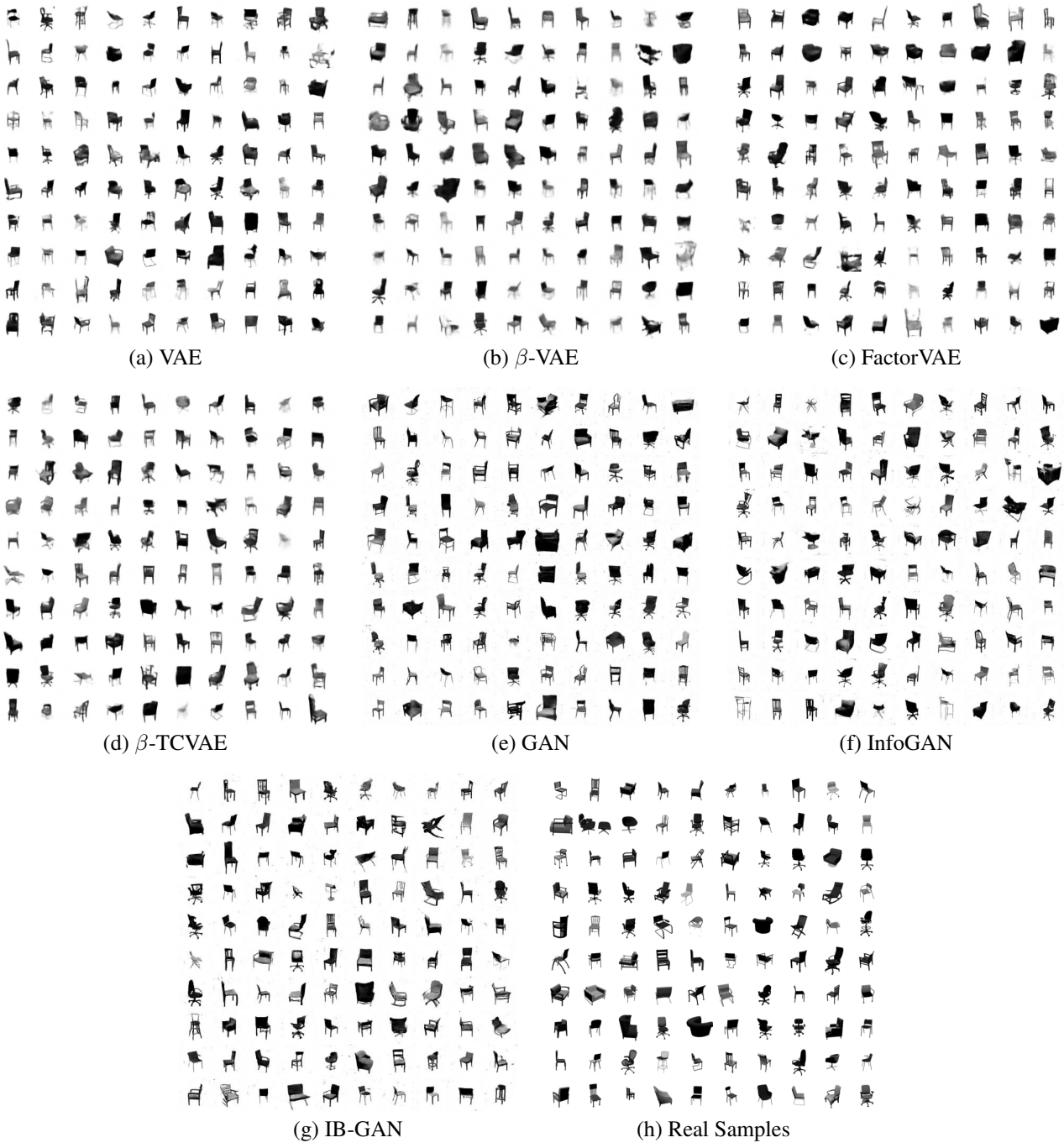
Figure 5 and 6 present randomly sampled images that are generated by the β-VAE and GAN baselines and IB-GAN on CelebA and 3D Chairs. For fair comparison of generation performance, we use the same architecture for the decoders of VAE baselines with the generator of IB-GAN. For the encoders of β-VAE baselines, we reverse the architecture of the decoder networks. The generated images by IB-GAN are often sharper and more realistic than those of β-VAE baselines (Kingma and Welling 2013; Higgins et al. 2017; Kim and Mnih 2018; Chen et al. 2018).

(a) VAE (b) β-VAE (c) FactorVAE

(d) β-TCVAE (e) GAN (f) InfoGAN

(g) IB-GAN (h) Real Samples

Figure 5: Comparison of randomly sampled images that are generated by the β-VAE and GAN baselines and IB-GAN on CelebA dataset.



(a) VAE (b) β-VAE (c) FactorVAE

(d) β-TCVAE (e) GAN (f) InfoGAN

(g) IB-GAN (h) Real Samples

Figure 6: Comparison of randomly sampled images that are generated by the β-VAE and GAN baselines and IB-GAN on 3D Chairs dataset.

C. Evaluation Metric

C.1 Disentanglement metric

We employ the metric proposed by Kim *et al*. (Kim and Mnih 2018) to evaluate the disentanglement performance of IB-GAN and other baselines. We use a batch of 100 samples to build each vote, and train a majority vote classifier using 800 votes. The accuracy of the classifier is then reported as the disentanglement performance for each model. Also, we exclude from consideration the collapsed latent dimensions of which empirical variances for the entire dataset are smaller than 0.05 (Kim and Mnih 2018; Locatello et al. 2018).

C.2 FID score

As commonly accepted, it is challenging to quantitatively evaluate how *good* generative models are. Nonetheless, the FID score (Heusel et al. 2017) is one possible candidate for measuring diversity and visual fidelity of generated samples. Precisely, the FID score measures the 2-Wasserstein distance between two distributions p and q: $F(q,p) = ||\mu_q - \mu_y||_2^2 + \text{trace}(C_q + C_p - 2(C_q C_p)^{1/2})$, where $\{\mu_q, C_q\}$ and $\{\mu_p, C_p\}$ are respectively the mean and the covariance of the feature vectors produced by the inception model (Szegedy et al. 2015) for true and generated samples. In both CelebA and 3D Chairs, we use 50,000 real and generated samples for the computation of C_q and C_p, respectively.

D. Model Architecture

Dataset	dSprites	Color-dSprites	CelebA	3DChairs
`ndf`	16	16	64	32
`ngf`	16	16	64	32

Table 1: The number of filters in the first layer of the generator (`ngf`) and discriminator and (`ndf`) for IB-GAN and GAN baselines.

Encoder (e)	Generator (G)	Discriminator (D) / Reconstructor (Q)
Input $z \in \mathbb{R}^{D_z}$	Input $r \in \mathbb{R}^{D_r}$	Input $x \in \mathbb{R}^{64\times64\times N_c}$
FC. 64 ReLU. BN	FC. `ngf`×16 ReLU. BN	4×4 conv. `ndf` lReLU. stride 2 (shared)
FC. 64 ReLU. BN	FC. $8 \times 8\times$ `ngf`×4 ReLU. BN	4×4 conv. `ndf`×2 lReLU. stride 2. BN (shared)
FC. 32 ReLU. BN	3×3 upconv. `ngf`×4 ReLU. stride 1. BN	4×4 conv. `ndf`×4 lReLU. stride 2. BN (shared)
FC $D_r \times 2$	3×3 upconv. `ngf`×4 ReLU. stride 1. BN	3×3 conv. `ndf`×4 lReLU. stride 1. BN
Reparametrization Trick	4×4 upconv. `ngf`×2 ReLU. stride 2. BN	3×3 conv. `ndf`×4 lReLU. stride 1. BN
–	4×4 upconv. `ngf`ReLU. stride 2. BN	8×8 conv. `ndf`×16 lReLU. stride 1. BN
–	4×4 upconv. N_c Tanh. stride 2	FC. D_z for Q FC. 1 for D

Table 2: The base architecture for IB-GAN on dSprites, CelebA, and 3DChairs. N_c denotes the number of channels of input images, lReLU is the leaky relu activation and BN is the batch normalization. The layers shared by discriminator D and reconstructor Q are marked as shared.

Encoder (e)	Generator (G)	Discriminator (D) / Reconstructor (Q)
Input $z \in \mathbb{R}^{D_z}$	Input $r \in \mathbb{R}^{D_r}$	Input $x \in \mathbb{R}^{64 \times 64 \times N_c}$
FC. 32 ReLU. BN	FC. ngf$\times$16 ReLU. BN	4×4 conv. ndf lReLU. stride 2
FC. 32 ReLU. BN	FC. $8 \times 8 \times$ ngf$\times$4 ReLU. BN	4×4 conv. ndf$\times$2 lReLU. stride 2. BN
FC $D_r \times 2$	3×3 upconv. ngf$\times$4 ReLU. stride 1. BN	4×4 conv. ndf$\times$4 lReLU. stride 2. BN
Reparametrization Trick	3×3 upconv. ngf$\times$4 ReLU. stride 1. BN	3×3 conv. ndf$\times$4 lReLU. stride 1. BN
–	4×4 upconv. ngf$\times$2 ReLU. stride 2. BN	3×3 conv. ndf$\times$4 lReLU. stride 1. BN
–	4×4 upconv. ngfReLU. stride 2. BN	8×8 conv. ndf$\times$16 lReLU. stride 1. BN
–	4×4 upconv. N_c Tanh. stride 2	FC. D_z for Q FC. 1 for D

Table 3: The base architecture for IB-GAN on Color-dSprites. N_c denotes the number of channels of input images, lReLU is the leaky relu activation and BN is the batch normalization.

Generator (G)	Discriminator (D) / Reconstructor (Q)
Input $z \in \mathbb{R}^{D_z}$	Input $x \in \mathbb{R}^{64 \times 64 \times N_c}$
FC. ngf$\times$16 ReLU. BN	4×4 conv. ndf lReLU. stride 2 (shared)
FC. $8 \times 8 \times$ ngf$\times$4 ReLU. BN	4×4 conv. ndf$\times$2 lReLU. stride 2. BN (shared)
3×3 upconv. ngf$\times$4 ReLU. stride 1. BN	4×4 conv. ndf$\times$4 lReLU. stride 2. BN (shared)
3×3 upconv. ngf$\times$4 ReLU. stride 1. BN	3×3 conv. ndf$\times$4 lReLU. stride 1. BN
4×4 upconv. ngf$\times$2 ReLU. stride 2. BN	3×3 conv. ndf$\times$4 lReLU. stride 1. BN
4×4 upconv. ngfReLU. stride 2. BN	8×8 conv. ndf$\times$16 lReLU. stride 1. BN
4×4 upconv. N_c Tanh. stride 2	FC. D_z for Q FC. 1 for D

Table 4: The base architecture for InfoGAN. This architecture is shared in the experiments on dSprites and Color-dSprites. N_c denotes the number of channels of input images, lReLU is the leaky relu activation and BN is the batch normalization. The layers shared by discriminator D and reconstructor Q are marked as shared. All hyperparameters are the same as those of IB-GAN except $D_z = 10$.

E. Implementation Details

Dataset	Hyperparameters	Learning rates	Iterations	Instance noise	Label smoothing
dSprites	D_z=16, D_r=10 γ=1, β=0.141	G/E/Q: 5e-5, D: 1e-6	1.5e5	$1 \rightarrow 0$ for 1e5 iters	No
Color-dSprites	D_z=16, D_r=10 γ=1, β=0.071	G/E/Q: 5e-5, D: 1e-6	5e5	$1 \rightarrow 0$ for 4e5 iters	No
3D Chairs	D_z=64, D_r=10 γ=1.2, β=0.35	G/E/Q: 5e-5, D: 2e-6	7e5	$0.5 \rightarrow 0.01$ for 7e5 iters	Yes
CelebA	D_z=64, D_r=15 γ=2, β=0.325	G/E/Q: 5e-5, D: 2e-6	1e6	$0.5 \rightarrow 0.01$ for 1e6 iters	Yes

Table 5: The hyperparameter settings for IB-GAN in all experiments. We use RMSProp with momention of 0.9. In hyperparameters, D_z and D_r mean the dimension of z and r. In learning rates, G,E,Q,D indicates generator, encoder, reconstructor and discriminator. In instance noise, σ_{inst} is annealed linearly between the two values for the following iterations.

Stabilization of GAN training. The training of GANs is notoriously unstable (Lučić et al. 2018; Mescheder, Geiger, and Nowozin 2018). To stabilize the training of GAN-based models in our experiments, we adopt two popular tricks: the instance noise technique (Sønderby et al. 2017) and one-sided label smoothing (Salimans et al. 2016). For the instance noise technique, we add instance noises $\epsilon \sim N(0, \sigma_{inst} * I)$ to both real and generated images while linearly decreasing the value of σ_{inst} during training iterations. For the one-sided label smoothing technique, we sample true labels from a uniform distribution within the range of [0.7, 1.2]. Table 5 summarizes important hyper-parameters including these two stabilization regularizers.

F. Datasets

Dataset	Specification
dSprites (Higgins et al. 2017)	737,280 binary 64×64 images of 2D shapes with 5 ground-truth factors, which consist of 3 shapes, 6 scales, 40 orientations and 32 positions of X and Y.
Color-dSprites (Burgess et al. 2018; Locatello et al. 2018)	RGB $64 \times 64 \times 3$ images of 2D shapes with 6 ground truth factors. All factors are identical to those of dSprites dataset, except for an additional *color* factor; it is quantized into 256 different bins, which are obtained by discretizing each color channel into 8 values linearly spaced between [0, 1].
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.

Table 6: The specification of datasets.

References

Alemi, A. A.; Fischer, I.; Dillon, J. V.; 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. 2018a. Fixing a Broken ELBO. In *ICLR*.

Alemi, A. A.; Poole, B.; Fischer, I.; Dillon, J.; Saurous, R. A.; and Murphy, K. 2018b. An information-theoretic analysis of deep latent-variable models. *https://openreview.net/forum?id=H1rRWl-Cb* .

Aubry, M.; Maturana, D.; Efros, A.; Russell, B.; and Sivic, J. 2014. Seeing 3D chairs: exemplar part-based 2D-3D alignment using a large dataset of CAD models. In *CVPR*.

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

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*.

Donahue, J.; Krähenbühl, P.; and Darrell, T. 2017. Adversarial Feature Learning. In *ICLR*.

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

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

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

Jordan, M. I.; Ghahramani, Z.; Jaakkola, T. S.; and Saul, L. K. 1999. An Introduction to Variational Methods for Graphical Models. *ML* .

Kim, H.; and Mnih, A. 2018. Disentangling by Factorising. In *ICML*.

Kingma, D. P.; and Welling, M. 2013. Auto-Encoding Variational Bayes. In *ICLR*.

Kulkarni, T. D.; Whitney, W. F.; Kohli, P.; and Tenenbaum, J. 2015. Deep convolutional inverse graphics network. In *NeurIPS*.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep Learning Face Attributes in the Wild. In *ICCV*.

Locatello, F.; Bauer, S.; Lucic, M.; Gelly, S.; Schölkopf, B.; and Bachem, O. 2018. Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations. *arXiv preprint arXiv:1811.12359* .

Lučić, M.; Kurach, K.; Michalski, M.; Gelly, S.; and Bousquet, O. 2018. Are GANs Created Equal? A Large-Scale Study. In *NeurIPS*.

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*.

Mescheder, L.; Geiger, A.; and Nowozin, S. 2018. Which Training Methods for GANs do actually Converge? In *ICML*.

Narayanaswamy, S.; Paige, T. 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*.

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

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

Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; and Chen, X. 2016. Improved techniques for training gans. In *NeurIPS*.

Sønderby, C. K.; Caballero, J.; Theis, L.; Shi, W.; and Huszár, F. 2017. Amortised MAP Inference for Image Super-resolution. In *ICLR*.

Springenberg, J. T. 2016. Unsupervised and semi-supervised learning with categorical generative adversarial networks. In *ICLR*.

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*.

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