High-Fidelity Synthesis with Disentangled Representation

Wonkwang Lee[1], Donggyun Kim[1], Seunghoon Hong[1](✉), and Honglak Lee[2,3]

[1] KAIST, Daejeon, South Korea
{wonkwang.lee,kdgyun425,seunghoon.hong}@kaist.ac.kr
[2] University of Michigan, Ann Arbor, USA
honglak@umich.edu
[3] Google AI, Cambridge, USA
honglak@google.com

Abstract. Learning disentangled representation of data without supervision is an important step towards improving the interpretability of generative models. Despite recent advances in disentangled representation learning, existing approaches often suffer from the trade-off between representation learning and generation performance (*i.e.,* improving generation quality sacrifices disentanglement performance). We propose an Information-Distillation Generative Adversarial Network (ID-GAN), a simple yet generic framework that easily incorporates the existing state-of-the-art models for both disentanglement learning and high-fidelity synthesis. Our method learns disentangled representation using VAE-based models, and distills the learned representation with an additional nuisance variable to the separate GAN-based generator for high-fidelity synthesis. To ensure that both generative models are aligned to render the same generative factors, we further constrain the GAN generator to maximize the mutual information between the learned latent code and the output. Despite the simplicity, we show that the proposed method is highly effective, achieving comparable image generation quality to the state-of-the-art methods using the disentangled representation. We also show that the proposed decomposition leads to an efficient and stable model design, and we demonstrate photo-realistic high-resolution image synthesis results (1024 × 1024 pixels) for the first time using the disentangled representations. Our code is available at https://www.github.com/1Konny/idgan.

1 Introduction

Learning a compact and interpretable representation of data without supervision is important to improve our understanding of data and machine learning systems. Recently, it is suggested that a *disentangled representation* that represents data

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-58574-7_10) contains supplementary material, which is available to authorized users.

A. Vedaldi et al. (Eds.): ECCV 2020, LNCS 12371, pp. 157–174, 2020.
https://doi.org/10.1007/978-3-030-58574-7_10

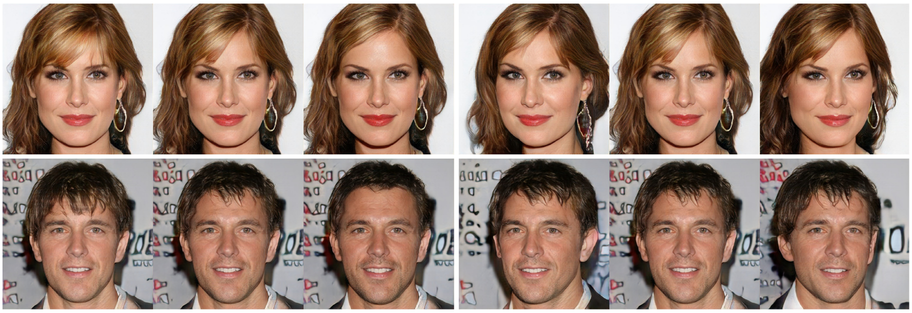

Fig. 1. Generated images on the CelebA-HQ dataset [35]. The proposed framework allows synthesizing high-resolution images (1024 × 1204 pixels) using the disentangled representation learned by VAEs.

using independent factors of variations in data can improve the interpretability and transferability of the representation [1,5,51]. Among various use-cases of disentangled representation, we are particularly interested in its application to generative models, since it allows users to specify the desired output properties by controlling the generative factors encoded in each latent dimension. There are increasing demands on such generative models in various domains, such as image manipulation [21,28,31], drug discovery [16], ML fairness [11,36], *etc.*(Fig. 1).

Most prior works on unsupervised disentangled representation learning formulate the problem as constrained generative modeling task. Based on well-established frameworks, such as the Variational Autoencoder (VAE) or the Generative Adversarial Network (GAN), they introduce additional regularization to encourage the axes of the latent manifold to align with independent generative factors in the data. Approaches based on VAE [7,9,18,26] augment its objective function to favor a factorized latent representation by adding implicit [7,18] or explicit penalties [9,26]. On the other hand, approaches based on GAN [10] propose to regularize the generator such that it increases the mutual information between the input latent code and its output.

One major challenge in the existing approaches is the trade-off between learning disentangled representations and generating realistic data. VAE-based approaches are effective in learning useful disentangled representations in various tasks, but their generation quality is generally worse than the state-of-the-arts, which limits its applicability to the task of realistic synthesis. On the other hand, GAN-based approaches can achieve the high-quality synthesis with a more expressive decoder and without explicit likelihood estimation [10]. However, they tend to learn comparably more entangled representations than the VAE counterparts [7,9,18,26] and are notoriously difficult to train, even with recent techniques to stabilize the training [26,54].

To circumvent this trade-off, we propose a simple and generic framework to combine the benefits of disentangled representation learning and high-fidelity synthesis. Unlike the previous approaches that address both problems jointly by a single objective, we formulate two separate, but successive problems; we first learn a disentangled representation using VAE, and *distill* the learned representation to GAN for high-fidelity synthesis. The distillation is performed from VAE to GAN by transferring the inference model, which provides a meaningful latent distribution, rather than a simple Gaussian prior and ensures that both models are aligned to render the same generative factors. Such decomposition also naturally allows a layered approach to learn latent representation by first learning major disentangled factors by VAE, then learning missing (entangled) nuisance factors by GAN. We refer the proposed method as the Information Distillation Generative Adversarial Network (ID-GAN).

Despite the simplicity, the proposed ID-GAN is extremely effective in addressing the previous challenges, achieving high-fidelity synthesis using the learned disentangled representation (*e.g.,* 1024 × 1024 image). We also show that such decomposition leads to a practically efficient model design, allowing the models to learn the disentangled representation from low-resolution images and transfer it to synthesize high-resolution images.

The contributions of this paper are as follows:

- We propose ID-GAN, a simple yet effective framework that combines the benefits of disentangled representation learning and high-fidelity synthesis.
- The decomposition of the two objectives enables plug-and-play-style adoption of state-of-the-art models for both tasks, and efficient training by learning models for disentanglement and synthesis using low- and high-resolution images, respectively.
- Extensive experimental results show that the proposed method achieves state-of-the-art results in both disentangled representation learning and synthesis over a wide range of tasks from synthetic to complex datasets.

2 Related Work

Disentanglement Learning. Unsupervised disentangled representation learning aims to discover a set of generative factors, whose element encodes unique and independent factors of variation in data. To this end, most prior works based on VAE [9,18,26] and GAN [10,22,33,34] focused on designing the loss function to encourage the factorization of the latent code. Despite some encouraging results, however, these approaches have been mostly evaluated on simple and low-resolution images [37,41]. We believe that improving the generation quality of disentanglement learning is important, since it not only increases the practical impact in real-world applications, but also helps us to better assess the disentanglement quality on complex and natural images where the quantitative evaluation is difficult. Although there are increasing recent efforts to improve the generation quality with disentanglement learning [22,33,34,45], they often come with the degraded disentanglement performance [10], rely on a specific inductive bias (*e.g.,* 3D transformation [45]), or are limited to low-resolution images [22,33,34]. On the contrary, our work aims to investigate a general framework to improve the generation quality without representation learning trade-off, while being general enough to incorporate various methods and inductive biases. We emphasize that this contribution is complementary to the recent efforts for designing better inductive bias or supervision for disentanglement learning [8,38,44,48,53]. In fact, our framework is applicable to a wide variety of disentanglement learning methods and can incorporate them in a plug-and-play style as long as they have an inference model (*e.g.,* nonlinear ICA [25]).

Combined VAE/GAN Models. There have been extensive attempts in literature toward building hybrid models of VAE and GAN [4,6,20,29,55], which learn to represent and synthesize data by jointly optimizing VAE and GAN objectives. Our method is an instantiation of this model family, but is differentiated from the prior work in that (1) the training of VAE and GAN is decomposed into two separate tasks and (2) the VAE is used to learn a specific conditioning variable (*i.e.,* disentangled representation) to the generator while the previous methods assume the availability of an additional conditioning variable [4] or use VAE to learn the entire (entangled) latent distribution [6,20,29,55]. Also, extending the previous VAE-GAN methods to incorporate disentanglement constraints is not straightforward, as the VAE and GAN objectives are tightly entangled in them. In the experiment, we demonstrate that applying existing hybrid models on our task suffers from the suboptimal trade-off between the generation and disentanglement performance, and they perform much worse than our method.

3 Background: Disentanglement Learning

The objective of unsupervised disentanglement learning is to describe each data x using a set of statistically independent generative factors z. In this section, we briefly review prior works and discuss their advantages and limitations.

The state-of-the-art approaches in unsupervised disentanglement learning are largely based on the Variational Autoencoder (VAE). They rewrite their original objective and derive regularizations that encourage the disentanglement of the latent variables. For instance, β-VAE [18] proposes to optimize the following modified Evidence Lower-Bound (ELBO) of the marginal log-likelihood:

$$\mathbb{E}_{x\sim p(x)}[\log p(x)] \geq \mathbb{E}_{x\sim p(x)}[\mathbb{E}_{z\sim q_\phi(z|x)}[\log p_\theta(x|z)] - \beta\ \mathrm{D}_{\mathrm{KL}}(q_\phi(z|x)||p(z))], \quad (1)$$

where setting $\beta = 1$ reduces to the original VAE. By forcing the variational posterior to be closer to the factorized prior ($\beta > 1$), the model learns a more disentangled representation, but with a sacrifice of generation quality, since it also decreases the mutual information between z and x [9,26]. To address such trade-off and improve the generation quality, recent approaches propose to gradually anneal the penalty on the KL-divergence [7], or decompose it to isolate the penalty for *total correlation* [52] that encourages the statistical independence of latent variables [1,9,26].

Approaches based on VAE have shown to be effective in learning disentangled representations over a range of tasks from synthetic [41] to complex datasets [3,35]. However, their generation performance is generally insufficient to achieve high-fidelity synthesis, even with recent techniques isolating the factorization of the latent variable [9,26]. We argue that this problem is fundamentally attributed to two reasons: First, most VAE-based approaches assume the fully-independent generative factors [9,18,26,37,40,51]. This strict assumption oversimplifies the latent manifold and may cause the loss of useful information (*e.g.,* correlated factors) for generating realistic data. Second, they typically utilize a simple generator, such as the factorized Gaussian decoder, and learn a uni-modal mapping from the latent to input space. Although this might be useful to learn meaningful representations [7] (*e.g.,* capturing a structure in local modes), such decoder makes it difficult to render complex patterns in outputs (*e.g.,* textures).

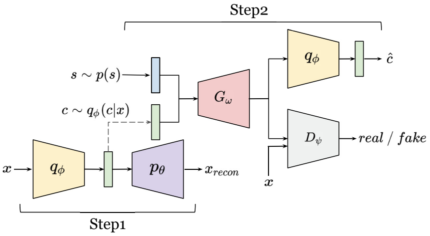

Fig. 2. Overall framework of the proposed method (ID-GAN).

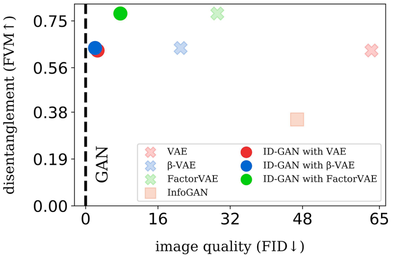

Fig. 3. Comparison of disentanglement vs. generation performance on dSprites dataset.

4 High-Fidelity Synthesis via Distillation

Our objective is to build a generative model $G_\omega : \mathcal{Z} \rightarrow \mathcal{X}$ that produces high-fidelity output $x \in \mathcal{X}$ with an interpretable latent code $z \in \mathcal{Z}$ (*i.e.,* disentangled representation). To achieve this goal, we build our framework upon VAE-based models due to their effectiveness in learning disentangled representations. However, discussions in the previous section suggest that disentanglement learning in VAE leads to the sacrifice of generation quality due to the strict constraints on fully-factorized latent variables and the utilization of simple decoders. We aim to improve the VAE-based models by enhancing generation quality while maintaining its disentanglement learning performance.

Our main idea is to decompose the objectives of learning disentangled representation and generating realistic outputs into separate but successive learning problems. Given a disentangled representation learned by VAEs, we train another network with a much higher modeling capacity (*e.g.,* GAN generator) to decode the learned representation to a realistic sample in the observation space.

Figure 2 describes the overall framework of the proposed algorithm. Formally, let $z = (s, c)$ denote the latent variable composed of the disentangled variable c

and the nuisance variable s capturing independent and correlated factors of variation, respectively. In the proposed framework, we first train VAE (*e.g.,* Eq. (1)) to learn disentangled latent representations of data, where each observation x can be projected to c by the learned encoder $q_\phi(c|x)$ after the training. Then in the second stage, we fix the encoder q_ϕ and train a generator $G_\omega(z) = G_\omega(s, c)$ for high-fidelity synthesis while *distilling* the learned disentanglement by optimizing the following objective:

$$\min_G \max_D \quad \mathcal{L}_{\text{GAN}}(D, G) - \lambda \mathcal{R}_{\text{ID}}(G), \tag{2}$$

$$\mathcal{L}_{\text{GAN}}(D, G) = \mathbb{E}_{x \sim p(x)}[\log D(x)] \ + \mathbb{E}_{s \sim p(s), c \sim q_\phi(c)}[\log(1 - D(G(s, c)))], \tag{3}$$

$$\mathcal{R}_{\text{ID}}(G) = \mathbb{E}_{c \sim q_\phi(c), x \sim G(s,c)}[\log q_\phi(c|x)] + H_{q_\phi}(c), \tag{4}$$

where $q_\phi(c) = \frac{1}{N}\sum_i q_\phi(c|x_i)$ is the aggregated posterior [19,39,50] of the encoder network[1]. Similar to [10], Eq. (4) corresponds to the variational lower-bound of mutual information between the latent code and the generator output $I(c; G(s, c))$, but differs in that (1) c is sampled from the aggregated posterior $q_\phi(c)$ instead of the prior $p(c)$ and (2) it is optimized with respect to the generator only. Note that we treat $H_{q_\phi}(c)$ as a constant since q_ϕ is fixed in Eq.(4). We refer the proposed model as the Information Distillation Generative Adversarial Network (ID-GAN).

4.1 Analysis

In this section, we provide in-depth analysis of the proposed method and its connections to prior works.

Comparisons to β-VAEs [9,18,26]. Despite the simplicity, the proposed ID-GAN effectively addresses the problems in β-VAEs with generating high-fidelity outputs; it augments the latent representation by introducing a nuisance variable s, which complements the disentangled variable c by modeling richer generative factors. For instance, the VAE objective tends to favor representational factors that characterize as much data as possible [7] (*e.g.,* azimuth, scale, lighting, *etc.*), which are beneficial in representation learning, but incomprehensive to model the complexity of observations. Given the disentangled factors discovered by VAEs, ID-GAN learns to encode the remaining generative factors (such as high-frequency textures, face identity, *etc.*) into nuisance variable s. (Fig. 8). This process shares a similar motivation with a progressive augmentation of latent factors [32], but is used for modeling disentangled and nuisance generative factors. In addition, ID-GAN employs a much more expressive generator than a simple factorized Gaussian decoder in VAE, which is trained with adversarial loss to render realistic and convincing outputs. Combining both, our method allows the generator to synthesize various data in a local neighborhood defined by c, where the specific characteristics of each example are fully characterized by the additional nuisance variable s.

[1] In practice, we can easily sample c from $q_\phi(c)$ by $c \sim q_\phi(c|x)p(x)$.

Comparisons to InfoGAN [10]. The proposed method is closely related to InfoGAN, which optimizes the variational lower-bound of mutual information $I(c; G(s,c))$ for disentanglement learning. To clarify the difference between the proposed method and InfoGAN, we rewrite the regularization for both methods using the KL divergence as follows:

$$\mathcal{R}_{\text{Info}}(G,q) = -\mathbb{E}_{s\sim p(s)}[D_{\text{KL}}(p(c)||q_\phi(c|G(s,c)))], \tag{5}$$

$$\mathcal{R}_{\text{ours}}(G,q) = \beta\mathcal{R}_{\text{VAE}}(q) + \lambda\mathcal{R}_{\text{ID}}(G), \text{ where}$$

$$\mathcal{R}_{\text{VAE}}(q) = -\mathbb{E}_{x\sim p(x)}[D_{\text{KL}}(q_\phi(c|x)||p(c))], \tag{6}$$

$$\mathcal{R}_{\text{ID}}(G) = -\mathbb{E}_{s\sim p(s)}[D_{\text{KL}}(q_\phi(c)||q_\phi(c|G(s,c)))], \tag{7}$$

where $\mathcal{R}_{\text{ours}}$ summarizes all regularization terms in our method[2]. See the Appendix A.1 for detailed derivations.

Equation (5) shows that InfoGAN optimizes the *forward* KL divergence between the prior $p(c)$ and the approximated posterior $q_\phi(c|G(s,c))$. Due to the zero-avoiding characteristics of forward KL [43], it forces all latent code c with non-zero prior to be covered by the posterior q_ϕ. Intuitively, it implies that InfoGAN tries to exploit every dimensions in c to encode each (unique) factor of variations. It becomes problematic when there is a mismatch between the number of true generative factors and the size of latent variable c, which is common in unsupervised disentanglement learning. On the contrary, VAE optimizes the *reverse* KL divergence (Eq. (6)), which can effectively avoid the problem by encoding only meaningful factors of variation into certain dimensions in c while collapsing the remainings to the prior. Since the encoder training in our method is only affected by Eq. (6), it allows us to discover the ambient dimension of latent generative factors robust to the choice of latent dimension $|c|$.

In addition, Eq. (5) shows that InfoGAN optimizes the encoder using the generated distributions, which can be problematic when there exists a sufficient discrepancy between the true and generated distributions (*e.g.,* mode-collapse may cause learning partial generative factors.). On the other hand, the encoder training in our method is guided by the true data (Eq. (6)) together with maximum likelihood objective, while the mutual information (Eq. (7)) is enforced only to the generator. This helps our model to discover comprehensive generative factors from data while guiding the generator to align its outputs to the learned representation.

Practical Benefits. The objective decomposition in the proposed method also offers a number of practical advantages. First, it enables plug-and-play-style adoption of the state-of-the-art models for disentangled representation learning and high-quality generation. As shown in Fig. 3, it allows our model to achieve state-of-the-art performance on both tasks. (Fig. 3). Second, such decomposition also leads to an efficient model design, where we learn disentanglement from

[2] In practice, we learn the encoder q_ϕ and generator G independently by Eq. (6) and (7), respectively, through two-step training.

low-resolution images and distill the learned representation to the task of high-resolution synthesis with a much higher-capacity generator. We argue that it is practically reasonable in many cases since VAEs tend to learn global structures in disentangled representation, which can be captured from low-resolution images. We demonstrate this in the high-resolution image synthesis task, where we use the disentangled representation learned with 64 × 64 images for the synthesis of 256 × 256 or 1024 × 1024 images.

5 Experiments

In this section, we present various results to show the effectiveness of ID-GAN. Please find the Appendix for more comprehensive results and figures.

Table 1. Quantitative comparison results on synthetic datasets.

	Color-dSprites			Scream-dSprites			Noisy-dSprites		
	FVM (↑)	MIG (↑)	FID (↓)	FVM (↑)	MIG (↑)	FID (↓)	FVM (↑)	MIG (↑)	FID (↓)
VAE [46]	.67±.12	.16±.08	21.63±4.97	.44±.03	.08±.04	7.79±2.51	**.42±.09**	.05±.04	3.27±1.94
β-VAE [18]	.67±.07	.32±.04	15.13±4.25	**.57±.01**	**.29±.00**	7.33±2.87	.32±.05	.05±.03	3.46±0.38
FactorVAE [26]	**.69±.05**	**.37±.02**	10.71±5.73	**.57±.01**	.22±.06	6.35±3.27	.40±.09	**.08±.04**	2.48±0.44
GAN [15]	N/A	N/A	.30±0.07	N/A	N/A	**.11±0.03**	N/A	N/A	9.74±2.18
InfoGAN [10]	.34± 00	.01±.01	30.55±21.17	.29± 00	.00±.00	5.77±3.93	.22±.02	.01±.01	5.51±4.22
OOGAN [34]	.32±.06	.01±.00	5.67±2.48	.21±.05	.00±.00	3.70±4.31	.21±.08	.01±.00	9.52±3.75
InfoGAN-CR [33]	.44±.10	.04±.02	.43±.19	.31±.08	.03±.05	6.05±5.28	.27±.02	.02±.01	48.52±57.12
ID-GAN+VAE	.67±.12	.16±.08	.32±0.10	.44±.03	.08±.04	.26±0.03	**.42±.09**	.05±.04	**1.58±0.62**
ID-GAN+β-VAE	.67±.07	.32±.04	**.25±0.23**	**.57±.01**	**.29±.00**	.18±0.02	.32±.05	.05±.03	12.42±1.13
ID-GAN+FactorVAE	**.69±.05**	**.37±.02**	.75±0.54	**.57±.01**	.22±.06	.65±0.33	.40±.09	**.08±.04**	2.07±0.87

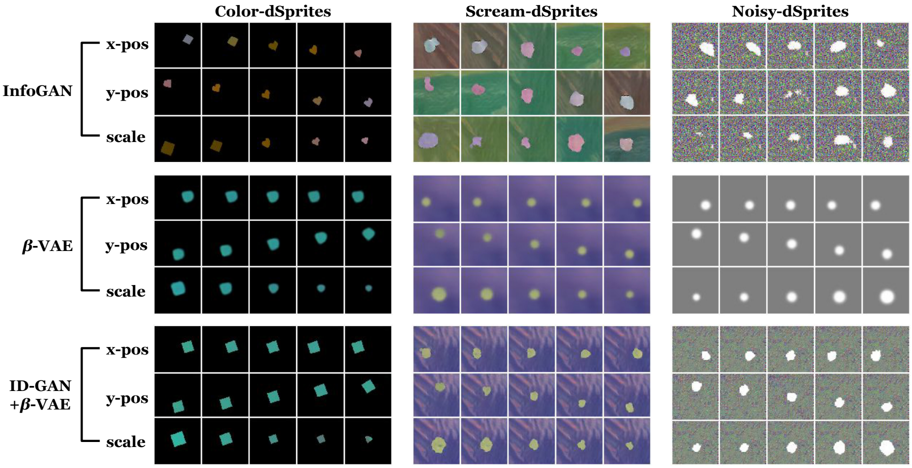

Fig. 4. Qualitative results on synthetic datasets. Both β-VAE and ID-GAN share the same latent code, but ID-GAN exhibits substantailly higher generation quality.

5.1 Implementation Details

Compared Methods. We compare our method with state-of-the-arts in disentanglement learning and generation. We choose β-VAE [18], FactorVAE [26], InfoGAN [10], OOGAN [34], and InfoGAN-CR [33] as baselines for disentanglement learning. For fair comparison, we choose the best hyperparameter for each model via extensive hyper-parameter search. We also report the performance by training each method over five different random seeds and averaging the results.

Network Architecture. For experiments on synthetic datasets, we adopt the architecture from [37] for all VAE-based methods (VAE, β-VAE, and FactorVAE). For GAN-based methods (GAN, InfoGAN, and ID-GAN), we employ the same decoder and encoder architectures in VAE as the generator and discriminator, respectively. We set the size of disentangled latent variable to 10 for all methods, and exclude the nuisance variable in GAN-based methods for a fair comparison with VAE-based methods. For experiments on complex datasets, we employ the generator and discriminator in the state-of-the-art GAN [42,47]. For VAE architectures, we utilize the same VAE architecture as in the synthetic datasets. We set the size of disentangled and nuisance variables to 20 and 256, respectively.

Evaluation Metrics. We employ three popular evaluation metrics in the literature: Factor-VAE Metric (FVM) [26], Mutual Information Gap (MIG) [9], and Fréchet Inception Distance (FID) [17]. *FVM* and *MIG* evaluate the disentanglement performance by measuring the degree of axis-alignment between each dimension of learned representations and ground-truth factors. *FID* evaluates the generation quality by measuring the distance between the true and the generated distributions.

5.2 Results on Synthetic Dataset

For quantitative evaluation of disentanglement, we employ the dSprites dataset [41], which contains synthetic images generated by randomly sampling known generative factors, such as shape, orientation, size, and x-y position. Due to the limited complexity of dSprites, we adopt three variants of dSprites, which are generated by adding color [26] (Color-dSprites) or background noise [37] (Noisy- and Scream-dSprites).

Table 1 and Fig. 4 summarize the quantitative and qualitative comparison results with existing disentanglement learning approaches, respectively. First, we observe that VAE-based approaches (*i.e.,* β-VAE and FactorVAE) achieve the state-of-the-art disentanglement performance across all datasets, outperforming the VAE baseline and InfoGAN with a non-trivial margin. The qualitative results in Fig. 4 show that the learned generative factors are well-correlated with meaningful disentanglement in the observation space. On the other hand, InfoGAN fails to discover meaningful disentanglement in most datasets. We observe that information maximization in InfoGAN often leads to undesirable factorization of generative factors, such as encoding both shape and position into one latent code, but factorizing latent dimensions by different combinations of them (*e.g.,* Color-dSprites in Fig. 4). ID-GAN achieves state-of-the-art disentanglement through the distillation of the learned latent code from the VAE-based models. Appendix B.3 also shows that ID-GAN is much more stable to train and insensitive to hyper-parameters than InfoGAN.

In terms of generation quality, VAE-based approaches generally perform much worse than GAN baseline. This performance gap is attributed to the strong constraints on the factorized latent variable and weak decoder in VAE, which limits the generation capacity. This is clearly observed in the results on the Noisy-dSprites dataset (Fig. 4), where the outputs from β-VAE fail to render the high-dimensional patterns in the data (*i.e.,* uniform noise). On the other hand, our method achieves competitive generation performance to the state-of-the-art GAN using a much more flexible generator for synthesis, which enables the modeling of complex patterns in data. As observed in Fig. 4, ID-GAN performs generation using the *same* latent code with β-VAE, but produces much more realistic outputs by capturing accurate object shapes (in Color-dSprites) and background patterns (in Scream-dSprites and Noisy-dSprites) missed by the VAE decoder. These results suggest that our method can achieve the best trade-off between disentanglement learning and high-fidelity synthesis.

Table 2. Comparison of approaches using a joint and decomposed objective for disentanglement learning and synthesis.

	dSprites		
	FVM (↑)	MIG (↑)	FID (↓)
β-VAE (reference)	**0.65±0.08**	**0.28±0.09**	37.75±24.58
VAE-GAN	0.46±0.18	0.13±0.11	33.54±24.93
ID-GAN (end-to-end)	0.50±0.14	0.13±0.09	3.18±2.38
ID-GAN (two-step)	**0.65±0.08**	**0.28±0.09**	**2.00±1.74**

5.3 Ablation Study

This section provides an in-depth analysis of our method.

Is Two-Step Training Necessary? First, we study the impact of two-stage training for representation learning and synthesis. We consider two baselines: (1) VAE-GAN [29] as an extension of β-VAE with adversarial loss, and (2) end-to-end training of ID-GAN. Contrary to ID-GAN that learns to represent (q_ϕ) and synthesize (G) data via separate objectives, these baselines learn a single, entangled objective for both tasks. Table 2 summarizes the results in the dSprites dataset.

The results show that VAE-GAN improves the generation quality of β-VAE with adversarial learning. The generation quality is further improved in the end-to-end version of ID-GAN by employing a separate generator for synthesis. However, the improved generation quality in both baselines comes with the cost of degraded disentanglement performance. We observe that updating the encoder using adversarial loss hinders the discovery of disentangled factors, as the discriminator tends to exploit high-frequency details to distinguish the real images from the fake images, which motivates the encoder to learn nuisance factors. This suggests that decomposing the representation learning and generation objective is important in the proposed framework (ID-GAN two-step), which achieves the best performance in both tasks.

Is Distillation Necessary? The above ablation study justifies the importance of two-step training. Next, we compare different approaches for two-step training that perform conditional generation using the representation learned by β-VAE.

Specifically, we consider two baselines: (1) cGAN and (2) ID-GAN trained without distillation (ID-GAN w/o distill). We opt to consider cGAN as the baseline since we find that it implicitly optimizes $\mathcal{R}_{\text{ID}}$ (see Appendix A.2 for the proof). In the experiments, we train all models in the CelebA 128 × 128 dataset using the same β-VAE trained on the 64 × 64 resolution, and compare the generation quality (FID) and a degree of alignment between the disentangled code c and generator output $G(s, c)$. For comparison of the alignment, we measure $\mathcal{R}_{\text{ID}}$ (Eq. (7)) and GILBO[3] [2], both of which are valid lower-bounds of mutual information $I(c; G(s, c))$. Note that the comparison based on the lower-bound is still valid as its relative order has shown to be insensitive to the tightness of the bound [2]. Table 3 and Fig. 5 summarize the quantitative and qualitative results, respectively.

Table 3. Comparison of two-step approaches for generation (FID) and alignment ($\mathcal{R}_{\text{ID}}$ and GILBO (We report both $\mathcal{R}_{\text{ID}}$ and GILBO without $H_{q_\phi}(c)$ to avoid potential error in measuring $q_\phi(c)$ (*e.g.,* fitting a Gaussian [2]). Note that it does not affect the relative comparison since all models share the same q_ϕ)) performance.

	CelebA 128 × 128		
	FID (↓)	$\mathcal{R}_{\text{ID}}$ (↑)	GILBO [2] (↑)
ID-GAN w/o distill	**5.75**	−65.84	−20.40
cGAN	7.07	−17.39	−7.57
ID-GAN	6.61	**−10.25**	**−0.19**

As shown in the table, all three models achieve comparable generation performances in terms of FID. However, we observe that their alignments to the input latent code vary across the methods. The qualitative results (Fig. 5) also show considerable mismatch between the c and the generated images. Compared

[3] GILBO is formulated similarly as $\mathcal{R}_{\text{ID}}$ (Eq. (4)), but optimized over another auxiliary encoder network different from the one used in $\mathcal{R}_{\text{ID}}$.

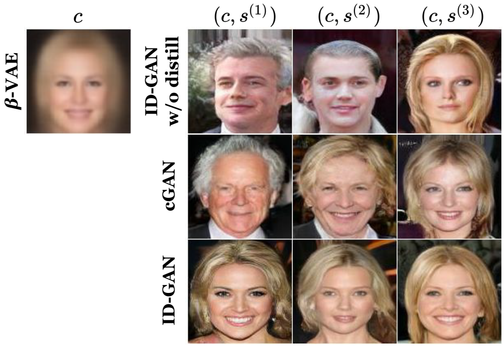

Fig. 5. Qualitative comparisons of various two-step approaches. All samples share the same disentangled code c, but different nuisance variable s. (1) First column: output of β-VAE decoder. (2) Second to fourth columns: images generated by different nuisance variables s using various methods (rows).

to this, cGAN achieves much higher degree of alignment due to the implicit optimization of $\mathcal{R}_{\text{ID}}$, but its association is much loose than our method (*e.g.*, changes in gender and hairstyle). By explicitly constraining the generator to optimize $\mathcal{R}_{\text{ID}}$, ID-GAN achieves the best alignment.

5.4 Results on Complex Dataset

To evaluate our method with more diverse and complex factors of variation, we conduct experiments on natural image datasets, such as CelebA [35], 3D Chairs [3], and Cars [27]. We first evaluate our method on 64 × 64 images, and extend it to higher resolution images using the CelebA (256 × 256) and CelebA-HQ [24] (1024 × 1024) datasets.

Table 4. Quantitative results based on FID (↓).

	3D chair	Cars	CelebA
VAE	116.46	201.29	160.06
βVAE	107.97	235.32	166.01
FactorVAE	123.64	208.60	154.48
GAN	**24.17**	14.62	**3.34**
InfoGAN	60.45	**13.67**	4.93
ID-GAN+βVAE	25.44	14.96	4.08

s

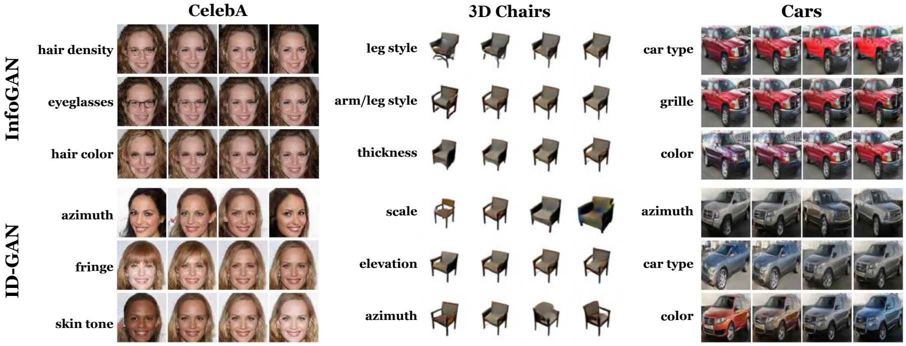

Fig. 6. Comparisons of latent traversal between GAN-based approaches. Despite the comparable generation quality, ID-GAN learns much more meaningful disentanglement.

Comparisons to Other Methods. Table 4 summarizes quantitative comparison results. Since the ground-truth factors are unknown, we report the performance based on generation quality (FID). As expected, the generation quality of VAEs is much worse in natural images. GAN-based methods, on the contrary, can generate more convincing samples although it tends to learn highly-entangled generative factors in nuisance variable. ID-GAN achieves disentanglement via disentangled factors learned by VAE, and generation performance on par with the GAN baseline. To better understand the disentanglement of GAN-based methods, we present latent traversal results in Fig. 6. We generate samples by modifying values of each dimension in the disentangled latent code c while fixing the rest. We observe that the InfoGAN fails to encode meaningful factors into c as the generation is dominated by the nuisance variable z, making all generated images almost identical. On the contrary, ID-GAN learns meaningful disentanglement with c and generates reasonable variations.

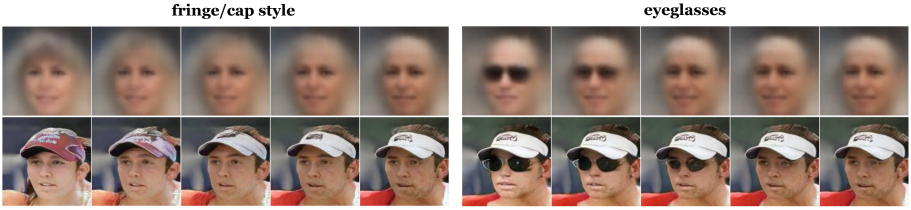

Fig. 7. Comparisons of VAE and ID-GAN outputs (top-rows: VAE, bottom-rows: ID-GAN). Note that both outputs are generated from the same latent code, but using different decoders. Both decoders are aligned well to render the same generative factors, but ID-GAN produces much more realistic outputs.

Extension to High-Resolution Synthesis. One practical benefit of the proposed two-step approach is that we can incorporate any VAE and GAN into our framework. To demonstrate this, we train ID-GAN for high-resolution images (*e.g.,* 256×256 and 1024×1024) while distilling the β-VAE encoder learned with *much smaller* 64×64 images[4]. This allows us to easily scale up the resolution of synthesis and helps us to better assess the disentangled factors.

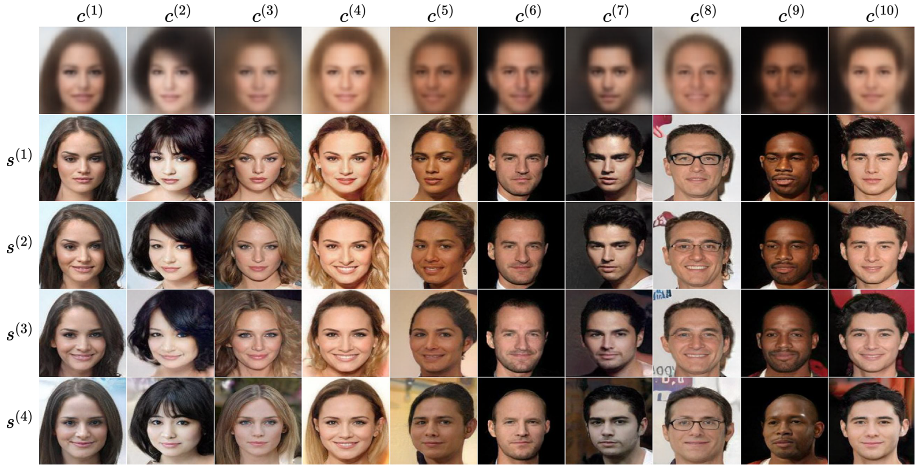

Fig. 8. Analysis on the learned disentangled variables $c^{(m)} \in \mathbb{R}^{20}$ and nuisance variables $s^{(n)} \in \mathbb{R}^{256}$ of ID-GAN on CelebA (256×256). The samples in the first row are generated by the β-VAE decoder and the rest are generated by ID-GAN. Each $c^{(m)}$ captures the most salient factors of variation (*e.g.,*azimuth) while $s^{(n)}$ contributes to the local details (*e.g.,*$s^{(2)}$ and $s^{(3)}$ for curvy and straight hair, respectively).

We first adapt ID-GAN to the 256×256 image synthesis task. To understand the impact of distillation, we visualize the outputs from the VAE decoder and the GAN generator using the same latent code as inputs. Figure 7 summarizes the results. We observe that the outputs from both networks are aligned well to render the same generative factors to similar outputs. Contrary to blurry and low-resolution (64×64) VAE outputs, however, ID-GAN produces much more realistic and convincing outputs by introducing a nuisance variable and employing more expressive decoder trained on higher-resolution (256×256). Interestingly, synthesized images by ID-GAN further clarify the disentangled factors learned by the VAE encoder. For instance, the first row in Fig. 7 shows that the ambiguous disentangled factors from the VAE decoder output is clarified by ID-GAN, which is turned out to capture the style of a cap. This suggests that ID-GAN can be useful in assessing the quality of the learned representation.

[4] We simply downsample the generator output by bilinear sampling to match the dimension between the generator and encoder.

To gain further insights on the learned generative factors by our method, we conduct qualitative analysis on the latent variables (c and s) by generating samples by fixing one variable while varying another (Fig. 8). We observe that varying the disentangled variable c leads to variations in the holistic structures in the outputs, such as azimuth, skin color, hair style, etc., while varying the nuisance variable s leads to changes in more fine-grained facial attributes, such as expression, skin texture, identity, *etc.*It shows that ID-GAN successfully distills meaningful and representative disentangled generative factors learned by the inference network in VAE, while producing diverse and high-fidelity outputs using generative factors encoded in the nuisance variable.

Finally, we further conduct experiments on the challenging task of mega-pixel image synthesis using CelebA-HQ dataset. We employ the generator architecture of VGAN [47] and adapt it to synthesize images given factors learned by β-VAE. Figure 9 presents the results, where we generate images by changing one values in one latent dimension in c. We observe that ID-GAN produces high-quality images with nice disentanglement, where it changes one factor of variation in the data (*e.g.,* azimuth and hair-style) while preserving the others (*e.g.,* identity).

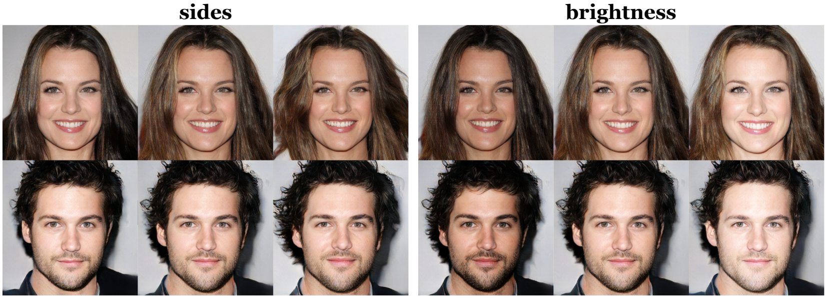

Fig. 9. Results on the CelebA-HQ dataset (1024 × 1024 images).

6 Conclusion

We propose Information Distillation Generative Adversarial Network (ID-GAN), a simple framework that combines the benefits of the disentanglement representation learning and high-fidelity synthesis. It allows us to incorporate the state-of-the-art for both tasks by decomposing their objectives while constraining the generator by distilling the encoder. Extensive experiments validate that the proposed method can achieve the best trade-off between realism and disentanglement, outperforming the existing approaches with substantial margin.

Acknowledgement. This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (2020-0-00153 and 2016-0-00464).

References

1. Achille, A., Soatto, S.: Information dropout: learning optimal representations through noisy computation. In: TPAMI (2018)
2. Alemi, A.A., Fischer, I.: GILBO: one metric to measure them all. In: NeurIPS (2018)
3. Aubry, M., Maturana, D., Efros, A., Russell, B., Sivic, J.: Seeing 3D chairs: exemplar part-based 2D–3D alignment using a large dataset of CAD models. In: CVPR (2014)
4. Bao, J., Chen, D., Wen, F., Li, H., Hua, G.: CVAE-GAN: fine-grained image generation through asymmetric training. In: ICCV (2017)
5. Bengio, Y., Courville, A., Vincent, P.: Representation Learning: a review and new perspectives. In: PAMI (2013)
6. Brock, A., Lim, T., Ritchie, J.M., Weston, N.: Neural photo editing with introspective adversarial networks. In: ICLR (2017)
7. Burgess, C.P., et al.: Understanding disentangling in β-VAE. In: NeurIPS (2017)
8. Chen, J., Batmanghelich, K.: Weakly supervised disentanglement by pairwise similarities. In: AAAI (2020)
9. Chen, T.Q., Li, X., Grosse, R., Duvenaud, D.: Isolating sources of disentanglement in variational autoencoders. In: NeurIPS (2018)
10. Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., Abbeel, P.: InfoGAN: interpretable representation learning by information maximizing generative adversarial nets. In: NeurIPS (2016)
11. Creager, E., et al.: Flexibly fair representation learning by disentanglement. In: ICML (2019)
12. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a Large-Scale Hierarchical Image Database. In: CVPR (2009)
13. Dosovitskiy, A., Brox, T.: Generating images with perceptual similarity metrics based on deep networks. In: NeurIPS (2016)
14. Fréchet, M.: Sur la distance de deux lois de probabilité. Comptes Rendus Hebdomadaires Des Seances de L'Academie Des Sciences (1957)
15. Goodfellow, I.J., et al.: Generative adversarial nets. In: NeurIPS (2014)
16. Gómez-Bombarelli, R., et al.: Automatic chemical design using a data-driven continuous representation of molecules. ACS Cent. Sci. **4**, 268–276 (2018)
17. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a nash equilibrium. In: NeurIPS (2017)
18. Higgins, I., et al.: β-VAE: learning basic visual concepts with a constrained variational framework. In: ICLR (2017)
19. Hoffman, M.D., Johnson, M.J.: ELBO surgery: yet another way to carve up the variational evidence lower bound. In: NeurIPS (2016)
20. Huang, H., Li, z., He, R., Sun, Z., Tan, T.: Introvae: introspective variational autoencoders for photographic image synthesis. In: NeurIPS. Curran Associates, Inc. (2018)
21. Huang, X., Liu, M.-Y., Belongie, S., Kautz, J.: Multimodal unsupervised image-to-image translation. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11207, pp. 179–196. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01219-9_11
22. Jeon, I., Lee, W., Kim, G.: IB-GAN: disentangled representation learning with information bottleneck GAN (2019). https://openreview.net/forum?id=ryljV2A5KX

23. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 694–711. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_43
24. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of GANS for improved quality, stability, and variation. In: ICLR (2018)
25. Khemakhem, I., Kingma, D., Hyvärinen, A.: Variational autoencoders and nonlinear ICA: a unifying framework. arXiv preprint arXiv:1907.04809 (2019)
26. Kim, H., Mnih, A.: Disentangling by factorising. In: ICML (2018)
27. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3D object representations for fine-grained categorization. In: 4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13) (2013)
28. Lample, G., Zeghidour, N., Usunier, N., Bordes, A., Denoyer, L., Ranzato, M.A.: Fader networks: manipulating images by sliding attributes. In: NeurIPS. Curran Associates, Inc. (2017)
29. Larsen, A.B.L., Sønderby, S.K., Larochelle, H., Winther, O.: Autoencoding beyond pixels using a learned similarity metric. In: ICML (2016)
30. Ledig, C., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: CVPR (2017)
31. Lee, H.-Y., Tseng, H.-Y., Huang, J.-B., Singh, M., Yang, M.-H.: Diverse image-to-image translation via disentangled representations. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11205, pp. 36–52. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01246-5_3
32. Lezama, J.: Overcoming the disentanglement vs reconstruction trade-off via Jacobian supervision. In: ICLR (2019)
33. d Lin, Z., Thekumparampil, K.K., Fanti, G.C., Oh, S.: InfoGAN-CR: disentangling generative adversarial networks with contrastive regularizers. In: ICML (2020)
34. Liu, B., Zhu, Y., Fu, Z., de Melo, G., Elgammal, A.: OOGAN: disentangling GAN with one-hot sampling and orthogonal regularization. In: AAAI (2020)
35. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: ICCV (2015)
36. Locatello, F., Abbati, G., Rainforth, T., Bauer, S., Schölkopf, B., Bachem, O.: On the fairness of disentangled representations. In: NeurIPS (2019)
37. Locatello, F., Bauer, S., Lucic, M., Gelly, S., Schölkopf, B., Bachem, O.: Challenging common assumptions in the unsupervised learning of disentangled representations. In: ICML (2019)
38. Locatello, F., Tschannen, M., Bauer, S., Rötsch, G., Schölkopf, B., Bachem, O.: Disentangling factors of variations using few labels. In: ICLR (2020)
39. Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., Frey, B.: Adversarial autoencoders. In: ICLR (2016)
40. Mathieu, E., Rainforth, T., Siddharth, N., Teh, Y.W.: Disentangling disentanglement in variational auto-encoders. In: Bayesian Deep Learning Workshop, NeurIPS (2018)
41. Matthey, L., Higgins, I., Hassabis, D., Lerchner, A.: dSprites: disentanglement testing sprites dataset (2017). https://github.com/deepmind/dsprites-dataset/
42. Mescheder, L., Nowozin, S., Geiger, A.: Which training methods for GANS do actually converge? In: ICML (2018)
43. Minka, T., et al.: Divergence measures and message passing. Technical report, Technical report, Microsoft Research (2005)
44. Narayanaswamy, S., et al.: Learning disentangled representations with semi-supervised deep generative models. In: NeurIPS (2017)

45. Nguyen-Phuoc, T., Li, C., Theis, L., Richardt, C., Yang, Y.L.: Hologan: unsupervised learning of 3D representations from natural images. In: ICCV (2019)
46. Kingma, D.P., Welling, M.: Auto-encoding variational Bayes. In: ICLR (2014)
47. Peng, X.B., Kanazawa, A., Toyer, S., Abbeel, P., Levine, S.: Variational discriminator bottleneck: improving imitation learning, inverse RL, and GANs by constraining information flow. In: ICLR (2019)
48. Ruiz, A., Martínez, O., Binefa, X., Verbeek, J.: Learning disentangled representations with reference-based variational autoencoders. arXiv preprint arXiv:1901.08534 (2019)
49. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: CVPR (2016)
50. Tolstikhin, I., Bousquet, O., Gelly, S., Schoelkopf, B.: Wasserstein auto-encoders. In: ICLR (2018)
51. Tschannen, M., Bachem, O.F., Lučić, M.: Recent advances in autoencoder-based representation learning. In: Bayesian Deep Learning Workshop, NeurIPS (2018)
52. Watanabe, S.: Information theoretical analysis of multivariate correlation. IBM J. Res. Dev. **4**, 66–82 (1960)
53. Watters, N., Matthey, L., Burgess, C.P., Lerchner, A.: Spatial Broadcast Decoder: a simple architecture for learning disentangled representations in VAEs. arXiv preprint arXiv:1901.07017 (2019)
54. Wei, X., Liu, Z., Wang, L., Gong, B.: Improving the improved training of Wasserstein GANs. In: ICLR (2018)
55. Zhu, J.Y., et al.: Toward multimodal image-to-image translation. In: NeurIPS (2017)