Semi-Supervised StyleGAN for Disentanglement Learning

Weili Nie [*1] **Tero Karras** [2] **Animesh Garg** [2 3] **Shoubhik Debnath** [2] **Anjul Patney** [2]
Ankit B. Patel [1] **Anima Anandkumar** [2 4]

Abstract

Disentanglement learning is crucial for obtaining disentangled representations and controllable generation. Current disentanglement methods face several inherent limitations: difficulty with high-resolution images, primarily focusing on learning disentangled representations, and non-identifiability due to the unsupervised setting. To alleviate these limitations, we design new architectures and loss functions based on StyleGAN (Karras et al., 2019), for semi-supervised high-resolution disentanglement learning. We create two complex high-resolution synthetic datasets for systematic testing. We investigate the impact of limited supervision and find that using only 0.25%∼2.5% of labeled data is sufficient for good disentanglement on both synthetic and real datasets. We propose new metrics to quantify generator controllability, and observe there may exist a crucial trade-off between disentangled representation learning and controllable generation. We also consider semantic fine-grained image editing to achieve better generalization to unseen images.

1. Introduction

Disentanglement learning with deep generative models has attracted much attention recently (Chen et al., 2016; Higgins et al., 2017; Locatello et al., 2019). This is crucial for controllable generation, where the style codes specified to the generator need to separately control various factors of variation for faithful generation. Another goal is learning disentangled representations where the input samples can be encoded to latent factors that are disentangled. This has been argued as a key to success of deep learning (Achille & Soatto, 2018). Previous works have primarily focused on only one of the above two objectives. Ideally, the ultimate goal of disentanglement learning is to achieve both the objectives at the same time, especially on more complex high-resolution images, and we pursue this goal in this paper. We first list the three main limitations of the current disentanglement methods.

First, much effort has focused on unsupervised disentanglement methods (Chen et al., 2016; Higgins et al., 2017; Nguyen-Phuoc et al., 2019). This is because a large number of fully annotated samples is expensive to obtain. These methods suffer from *non-identifiability*, which means multiple repeated runs will not reliably observe the same latent representations (Hyvärinen & Pajunen, 1999; Locatello et al., 2019). In addition, human feedback is needed to discern (i) what factors of variation the model has learnt (e.g., object shape and color), and (ii) what semantic meaning different values of the discovered factor code represent (e.g., red and blue in a color factor code). To reliably control generation for practical use, adding a small amount of labeled data may resolve the non-identifiability issue and lead to interpretable factors. Hence, we investigate the impact of *limited supervision* on disentanglement learning.

Second, current disentanglement methods (Locatello et al., 2019; 2020) are mainly developed and evaluated on relatively simple low-resolution images, such as dSprites (Matthey et al., 2017) and 3DShapes (Kim & Mnih, 2018), which raises concerns about their ability to scale up to more diverse, higher-resolution images. For example, the use of 3D representations to disentangle the 3D pose may not easily apply to high-resolution images due to the computational cost. The difficulty of some deep generative models at generating realistic images also limits their application in more complex domains. Furthermore, although there exist many real image datasets of high resolution, the latent factors are typically only partially observed or unbalanced, which makes it hard to scientifically study disentanglement. To gain practically useful insights, it is critical to first test disentanglement methods on complex, high-resolution, synthetic images wherein ground-truth factors are easy to obtain.

Third, most previous works (Kim & Mnih, 2018; Chen et al., 2018) have primarily focused on learning disentangled representations by quantifying encoder disentanglement quality, in the hopes that a better disentangled encoder might also

[1]Rice University ([*]Work done as a part of internship at Nvidia) [2]Nvidia [3]University of Toronto [4]California Institute of Technology. Correspondence to: Weili Nie <wn8@rice.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

lead to a better disentangled generator. However, to the best of our knowledge, there is no clear evidence supporting that a proportional relationship between encoder and generator disentanglement quality always exists. A good analogy is that an art critic can disentangle various painting styles and skills, but may not be able to create a good painting by combining these styles and skills. Thus, results based solely on evaluating encoder disentanglement may be misleading, especially in tasks where the requirement for controllable generation is more critical. This highlights the importance of measuring the generator disentanglement quality in order to properly evaluate and compare different methods.

Main contributions. In this work, we investigate semi-supervised disentanglement learning based on StyleGAN (Karras et al., 2019), one of the state-of-the-art generative adversarial networks (GANs), for complex high-resolution images. In summary, our main contributions are as follows,

- We first justify the advantages of the StyleGAN architecture in disentanglement learning, by showing that StyleGAN augmented with a mutual information loss (called Info-StyleGAN) ourperforms most state-of-the-art unsupervised disentanglement methods.
- We propose Semi-StyleGAN that achieves near fully-supervised disentanglement quality with limited supervision (0.25%$\sim$2.5%) on synthetic and real data.
- We propose new metrics (termed as MIG-gen and L2-gen) to evaluate the generator controllability, and reveal a crucial trade-off between learning disentangled representations and controllable generation.
- We then extend Semi-StyleGAN to an image-to-image model, enabling semantic fine-grained image editing with better generalization to unseen images.
- We create two high-quality datasets with much higher resolution, better photorealism, and richer factors of variation than existing disentanglement datasets.

2. Background and Related Work

StyleGAN. GANs are a family of generative models that have shown great success. Among various GANs, StyleGAN (Karras et al., 2019) is a state-of-the-art GAN architecture for unsupervised image generation, particularly for high-fidelity human faces of resolution up to 1024x1024. StyleGAN comprises a mapping network whose role is to map a latent vector z to an intermediate space, which then controls the styles at each convolutional layer in the synthesis network with adaptive instance normalization (AdaIN) (Ulyanov et al., 2016; Huang & Belongie, 2017). StyleGAN also enables the separation of fine-grained and coarse-grained features. For example, modifying the styles of low-resolution blocks affects only coarse-grained features (e.g. overall pose and presence of eyeglasses), while modifying the styles of high-resolution blocks affects only fine-grained features (e.g. color scheme and microstructure). These nice properties make it a potentially good candidate for disentanglement learning of high-resolution images.

Disentanglement learning. In terms of unsupervised disentanglement learning, there exists much prior work based on either Variational Autoencoders (VAEs) (Higgins et al., 2017; Kim & Mnih, 2018; Chen et al., 2018) or GANs (Chen et al., 2016; Lin et al., 2019; Nguyen-Phuoc et al., 2019). The basic idea in disentangled VAEs is to encourage a factorization of the latent code by regularizing the *total correlation* (Chen et al., 2018). They have shown the state-of-the-art performance on many standard disentanglement benchmarks (Matthey et al., 2017; Kim & Mnih, 2018). Many GAN-based models reply on maximizing the mutual information between the observation and factor code, such as InfoGAN (Chen et al., 2016) and its variants (Lin et al., 2019). Other GANs also learn disentanglement by designing a domain-specific generator architecture to add model-inductive bias, represented by HoloGAN (Nguyen-Phuoc et al., 2019). However, the use of 3D representations in HoloGAN may not scale up to higher-resolution images.

Another line of work in disentanglement learning is to use explicit supervision. (Kulkarni et al., 2015) applies a supervised training procedure to encourage each group of the graphics code to distinctly represent a specific factor of variation. (Bouchacourt et al., 2018) proposes the Multi-Level VAE (ML-VAE) to learn disentanglement from the supervision of group information. (Xiao et al., 2017) develops a supervised disentanglement algorithm called DNA-GAN using a swapping policy. (Narayanaswamy et al., 2017) proposes a semi-supervised VAE by employing a general graphical model structure in the encoder and decoder. However, it still remains unclear how the use of supervision impacts the disentanglement learning. (Spurr et al., 2017) proposes ss-InfoGAN by adding few labels to InfoGAN to learn semantically meaningful data representations. More recently, (Locatello et al., 2020) shows the benefits of adding limited supervision into learning disentangled presentations in VAEs. We extend these results to GANs on more complex and higher-resolution images, and also quantify the impact of limited supervision on the generator controllability.

Conditional GANs. A class of conditional GANs conditions on the class labels for better image generation quality, such as cGAN (Mirza & Osindero, 2014), AC-GAN (Odena et al., 2017), Projection Discriminator (Miyato & Koyama, 2018). The architectural component of semi-StyleGAN makes it a special case of semi-supervised conditional GANs, as it conditions on partially available factor codes. However, the tasks and loss functions of Semi-StyleGAN differ greatly from those of these conditional GANs. First, these conditional GANs mainly focus on generating more

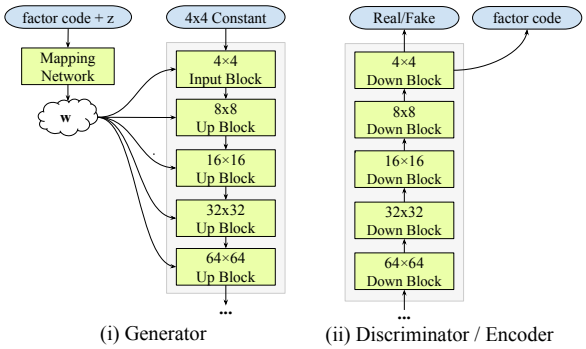


Figure 1. An illustration of disentanglement learning based on StyleGAN, where the mapping network in the generator conditions on the factor code and the encoder (which shares all layers in the discriminator except for the last layer) predicts its value.

realistic images, whereas semi-StyleGAN focuses on two joint tasks: (i) disentangled representation learning and (ii) controllable generation. Second, to this end, both the mutual information loss and a new smoothness regularization are introduced in our loss functions for a better trade-off in disentanglement learning.

Another class of conditional GANs conditions on the given images for image-to-image translation, where many works focus on multi-attribute editing, and two representatives are StarGAN (Choi et al., 2018) and AttGAN (He et al., 2019). Although the proposed Semi-StyleGAN-*fine* in Section 5 share the similar objectives, there are many differences: StarGAN needs to condition on images of a specific domain, namely a set of images sharing the same attribute, which limits itself to only performing discrete/binary attribute control. However, our work is capable of doing continuous style manipulation. AttGAN requires an encoder-decoder structure in the generator while we do not need to. Instead, our work adds controllable fine-grained factors along with a super-resolution process, which is especially suitable for editing fine styles of high-resolution images with a good generalization ability.

3. Why StyleGAN for Disentanglement?

In unsupervised disentanglement learning, there exist many state-of-the-art VAE-based models, such as β-VAE (Higgins et al., 2017), FactorVAE (Kim & Mnih, 2018) and β-TCVAE (Chen et al., 2018). More recently, GAN-based models, such as InfoGAN-CR (Lin et al., 2019), have also achieved competitive performance by adding more tuning heuristics and regularization. Here, we consider Info-StyleGAN, enabling StyleGAN with a mutual information loss, and show that the structural advance of StyleGAN provides a stronger prior for disentanglement learning compared to regularization used previously in VAEs or GANs.

As shown in Figure 1, the mapping network in the generator of Info-StyleGAN now conditions on a *factor code*, a vector representing each factor of variation in each dimension, by simply concatenating it with the latent code z. The output of the mapping network, called *conditional styles* will modulate each block in the synthesis network using AdaIN. Similar to InfoGAN, the encoder in Info-StyleGAN shares all the network layers except the last one with the discriminator and predicts the value of factor code. Thus, we use $G/D/E$ to represent the generator, discriminator and encoder, respectively. The mutual information loss of InfoGAN can be approximated by an *unsupervised code reconstruction* loss (Chen et al., 2016), which is

$$\mathcal{L}_{\text{unsup}} = \sum\nolimits_{c\sim\mathcal{C}, z\sim p_z} \|E(G(c,z)) - c\|_2 \tag{1}$$

where $\mathcal{C}$ denotes the set of all factor codes, and p_z denote the prior distribution of latent code z. The respective loss functions for G and (D, E) are given by

$$\begin{aligned} \mathcal{L}^{(G)} &= \mathcal{L}_{\text{GAN}} + \gamma\mathcal{L}_{\text{unsup}} \\ \mathcal{L}^{(D,E)} &= -\mathcal{L}_{\text{GAN}} + \gamma\mathcal{L}_{\text{unsup}} \end{aligned} \tag{2}$$

where we keep the GAN loss function $\mathcal{L}_{\text{GAN}}$ the same as in (Karras et al., 2019). The hyperparameter γ controls a trade-off between image realism and disentanglement quality.

3.1. Experimental Setup

Datasets and evaluation metrics. We consider two datasets to compare Info-StyleGAN and state-of-the-art disentanglement models: dSprites (Matthey et al., 2017) and our proposed Isaac3D (See details in Section 4.1). dSprites is a commonly used dataset in disentanglement learning, with 737,280 images and each of resolution 64x64. For experiments on dSprites, we use both *Factor score* (Kim & Mnih, 2018) and *Mutual Information Gap (MIG)* (Chen et al., 2018) to evaluate disentanglement. For experiments on Isaac3D, we first downscale the resolution of each image to 128x128 because VAEs have difficulties in generating higher-resolution images. We use MIG to evaluate disentanglement quality and *Frechet Inception Distance (FID)* (Heusel et al., 2017) to evaluate image quality.

Experimental protocol. We consider β-VAE, FactorVAE, β-TCVAE and InfoGAN-CR for comparison, where all the models are trained based on the implementation in (Locatello et al., 2019). For VAEs, we set $\beta = 6$ for β-VAE, $\gamma = 30$ for FactorVAE and $\beta = 8$ for β-TCVAE after a grid search over different hyperparameters. For InfoGAN-CR, we use default hyperparameters on dSprites from the original paper, and perform a grid search over different hyperparameters on Isaac3D to report the best results. For Info-StyleGAN, we keep $\gamma = 10$ on dSprites and $\gamma = 1$ on Isaac3D. We also keep the progressive training (Karras et al., 2017), as we show that it helps improve disentanglement

Methods	# Params	Factor Score ↑	MIG ↑
β-VAE	0.69M	0.713 ± 0.095	0.132 ± 0.031
FactorVAE	5.70M	0.764 ± 0.098	0.175 ± 0.057
β-TCVAE	0.69M	0.731 ± 0.097	0.174 ± 0.046
InfoGAN-CR	0.76M	$\mathbf{0.853 \pm 0.046}$	0.270 ± 0.034
Info-StyleGAN*	0.74M	0.769 ± 0.144	0.274 ± 0.096
Info-StyleGAN	47.89M	0.840 ± 0.090	$\mathbf{0.290 \pm 0.098}$

(a) dSprites with resolution 64x64

Methods	# Params	FID ↓	MIG ↑
β-VAE	1.91M	122.6 ± 2.0	0.231 ± 0.068
FactorVAE	6.93M	305.8 ± 142.1	0.245 ± 0.034
β-TCVAE	1.91M	155.4 ± 13.6	0.216 ± 0.074
InfoGAN-CR	3.29M	80.72 ± 30.79	0.342 ± 0.139
Info-StyleGAN*	3.44M	8.10 ± 2.25	$\mathbf{0.404 \pm 0.085}$
Info-StyleGAN	49.05M	$\mathbf{2.19 \pm 0.48}$	0.328 ± 0.057

(b) Isaac3D with resolution 128x128

Table 1. Comparison of Info-StyleGAN and state-of-the-art disentanglement models on dSprites and (downscaled) Isaac3D. Note that the scores of VAEs are obtained based on the implementation in (Locatello et al., 2019). Info-StyleGAN* represents the smaller version of Info-StyleGAN, in which its number of parameters (i.e., # Params) is similar to that of previous models.

in Appendix B.1. Because Info-StyleGAN and state-of-the-art disentanglement models have largely different network architectures, for a fairer comparison, we also try to keep their network sizes to be the same. See Appendix B.2 for how we decrease the network size of Info-StyleGAN (called Info-StyleGAN*) to match those of previous models.

3.2. Key Results

Table 1 shows that Info-StyleGAN and its variant with smaller network size, termed as Info-StyleGAN*, consistently outperform state-of-the-art VAE-based methods by a large margin on both dSprites and Isaac3D. Meanwhile, Info-StyleGAN achieves competitive or even better disentanglement performance than the strong GAN baseline. Although unsupervised disentanglement learning is impossible without supervision or inductive bias (Locatello et al., 2019), this result reveals that the network structural improvement of StyleGAN provides *a stronger prior* for disentanglement learning compared to different explicit loss regularizations in disentangled VAEs or InfoGAN-CR. Besides, we observe that previous methods have much higher FID scores on (downscaled) Isaac3D, along with their poor generated samples in Appendix B.2. We have also increased the capacity of VAEs but the improvement of image quality still cannot close the gap with Info-StyleGAN, as shown in Appendix B.4. These results show that previous disentanglement methods have difficulties on more diverse and complex data, such as Isaac3D, while StyleGAN does not.

4. Semi-StyleGAN

As pointed out by (Locatello et al., 2019) that unsupervised disentangled methods are formally non-identifiable (Hyvärinen & Pajunen, 1999), the impact of *limited supervision* on both learning disentangled representations and controllable generation, which has been rarely explored, becomes crucial. In this section, we propose to add semi-supervision into Info-StyleGAN to get a semi-supervised disentanglement model – *Semi-StyleGAN*. Based on Semi-StyleGAN, we systematically analyze the role of limited supervision on both synthetic and real data.

A naive way of applying (semi-)supervision is to add a *supervised code reconstruction* term for the small amount of labeled data into Eq (2), similar to (Kingma et al., 2014; Odena et al., 2017; Locatello et al., 2020). That is,

$$\mathcal{L}_{\text{sup}} = \sum\nolimits_{(x,c)\sim\mathcal{J}} \|E(x) - c\|_2 \tag{3}$$

where $\mathcal{J}$ represents the set of labeled pairs of real image and factor code. When considering the *limited supervision*, we assume the cardinality of the labeled set $\mathcal{J}$ satisfies that $|\mathcal{J}| \ll |\mathcal{X}|$, with $\mathcal{X}$ being the set of all real images. Thus, the semi-supervised loss functions become

$$\begin{aligned} \mathcal{L}^{(G)} =& \mathcal{L}_{\text{GAN}} + \gamma_G \mathcal{L}_{\text{unsup}} \\ \mathcal{L}^{(D,E)} =& -\mathcal{L}_{\text{GAN}} + \gamma_E \mathcal{L}_{\text{unsup}} + \beta \mathcal{L}_{\text{sup}} \end{aligned} \tag{4}$$

where β is the weight of the supervised term $\mathcal{L}_{\text{sup}}$, and we use different γ's (denoted by γ_G and γ_E) to separately represent the weight of the unsupervised term in $\mathcal{L}^{(G)}$ and $\mathcal{L}^{(D,E)}$. As we show later, γ_G and γ_E play an important role in controlling the trade-off between encoder and generator disentanglement. Note that the supervised term $\mathcal{L}_{\text{sup}}$ does not update G directly as shown in Eq. (4).

While semi-supervised learning for image recognition is an active research area, many algorithms may not be directly applied to disentanglement learning. Take the consistency regularization (Sajjadi et al., 2016) as an example. Commonly used data perturbations, such as image rotation and color randomization will inevitably cause inconsistency if the considered factors of variation include object rotation or color. In contrast, encouraging smoothness in the latent space of GANs may help improve disentanglement (Karras et al., 2019). Thus, we propose to explicitly add a smoothness regularization by using the idea of MixUp (Zhang et al., 2018; Berthelot et al., 2019).

Formally, given a labeled observation-code pair $(x, c) \sim \mathcal{J}$ and a generated pair (x', c') where $x' = G(z, c')$, we get a set of mixed observation-code pairs $\mathcal{M} = \{(\tilde{x}, \tilde{c})\}$ by

$$\begin{aligned} \lambda \sim& \text{Beta}(\xi, \xi), \quad \lambda' = \max(\lambda, 1-\lambda) \\ \tilde{x} =& \lambda' x + (1-\lambda') x' \\ \tilde{c} =& \lambda' c + (1-\lambda') c' \end{aligned} \tag{5}$$

where ξ is a hyperparameter. Thus, the smoothness regularization term is

$$\mathcal{L}_{\text{sr}} = \sum\nolimits_{(x,c)\sim\mathcal{M}} \|E(x) - c\|_2 \tag{6}$$

and the new semi-supervised loss functions with smoothness regularization become

$$\begin{aligned}\mathcal{L}^{(G)} =& \mathcal{L}_{\text{GAN}} + \gamma_G \mathcal{L}_{\text{unsup}} + \alpha \mathcal{L}_{\text{sr}} \\ \mathcal{L}^{(D,E)} =& -\mathcal{L}_{\text{GAN}} + \gamma_E \mathcal{L}_{\text{unsup}} + \beta \mathcal{L}_{\text{sup}} + \alpha \mathcal{L}_{\text{sr}}\end{aligned} \tag{7}$$

where α is the weight of the smoothness term $\mathcal{L}_{\text{sr}}$. Different from (Zhang et al., 2018; Berthelot et al., 2019) that combines labeled and unlabeled real data, the MixUp in (5) is performed between real labeled data and generated data. This way, it not only encourages smooth behaviors of both the generator and encoder, but also takes good advantages of enormous fake data for disentanglement.

4.1. New Datasets

Current disentanglement datasets, such as dSprites (Matthey et al., 2017), 3DShapes (Kim & Mnih, 2018) and MPI3D (Gondal et al., 2019), are of low resolution and mostly lack photorealism. We create two new datasets – *Falcor3D* and *Isaac3D*, with much higher resolution, better photorealism and richer factors of variations, as shown in Table 2.

Falcor3D. It contains 233,280 images and each has a resolution of 1024x1024. This dataset is based on the 3D scene of a living room, where we move the camera positions and change the lighting conditions. Each image is paired with a ground-truth factor code, consisting of 7 factors of variation: lighting intensity (5), lighting x-dir (6), lighting y-dir (6), lighting z-dir (6), camera x-pos (6), camera y-pos (6), and camera z-pos (6). The number m behind each factor represents that the factor has m possible values, uniformly sampled from $[0, 1]$. For example, "lighting x-dir (6)" represents the lighting direction moving along the x-axis and "camera z-pos (6)" denotes the camera position moving along the z-axis. Both factors have 6 possible values.

Isaac3D. It contains 737,280 images and each has a resolution of 512x512. This dataset is based on the 3D scene of a kitchen, where we move the camera positions and vary the lighting conditions. There is a robotic arm inside, grasping an object. The robotic arm has two degrees of freedom: x-movement (horizontal) and y-movement (vertical). The attached object could change its shape, scale or color. All objects in the 3D scene are properly textured for better photorealism. Similarly, each image is paired with a ground-truth factor code, consisting of 9 factors of variation: lighting intensity (4), lighting y-dir (6), object color (4), wall color (4), object shape (3), object scale (4), camera height (4), robot x-movement (8), and robot y-movement (5). The number m behind each factor represents that it has m possible values, uniformly sampled from $[0, 1]$.

4.2. New Metrics

Many metrics have been proposed for evaluating disentanglement, such as Factor score (Kim & Mnih, 2018), MIG (Chen et al., 2018), DCI score (Ridgeway & Mozer, 2018), and SAP score (Kumar et al., 2017). See the prior work (Locatello et al., 2019) for their more implementation details. However, they all have some inherent limitations in quantifying semi-supervised disentanglement methods.

Datasets	# Images	# Factors	Resolution	3D
dSprites	737,280	5	64x64	✗
Noisy dSprites	737,280	7	64x64	✗
Scream dSprites	737,280	7	64x64	✗
SmallNORB	48,600	5	128x128	✓
Cars3D	17,568	3	64x64	✓
3DShapes	480,000	7	64x64	✓
MPI3D	640,800	7	64x64	✓
Falcor3D	233,280	7	1024x1024	✓
Isaac3D	737,280	9	512x512	✓

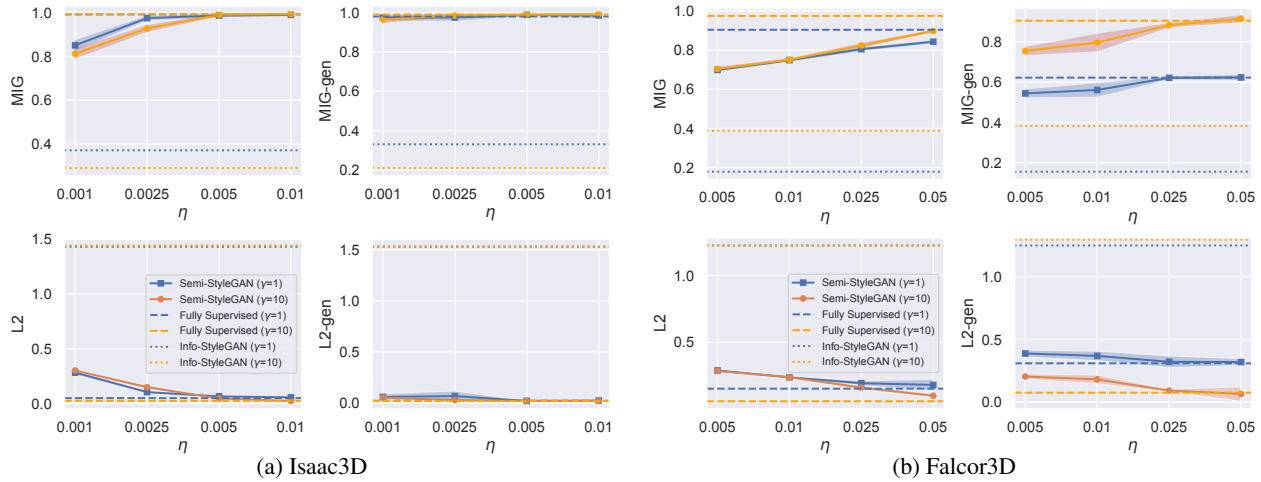
Table 2. Summary of the proposed two datasets, compared with currently commonly used datasets (Gondal et al., 2019). We can see that the proposed two datasets – *Faclor3D* and *Isaac3D* have much larger resolutions than previous datasets, along with the largest number of factors. More importantly, the proposed datasets are of much higher photorealism, as shown in Appendix A.

First, these metrics are designed for *unsupervised* disentanglement methods, which are non-identifiable (Locatello et al., 2019). But with supervision, the model become identifiable and thus we need to evaluate the semantic meaning of learned representations as well. A simple solution here is to measure the average $L2$ distance between the ground-truth factor code and the prediction of its paired observation using the considered encoder, termed as $L2$ score.

Second, these metrics only evaluate the the encoder disentanglement while ignoring the generator controllability, another important characteristic of disentanglement learning. However, there may exist a trade-off between the encoder and generator disentanglement. That is, a high MIG score does not mean a good model in terms of the controllable generation ability. Therefore, we propose new metrics to quantify the generator controllability.

Specifically, given a generator G to be evaluated and an oracle encoder E_{oracle} that can perfectly predict the factor code, we first sample N generated observation-code pairs $(x'^{(n)}, c'^{(n)})$ where $x'^{(n)} = G(z, c'^{(n)})$. We then pass the generated sample $x'^{(n)}$ into E_{oracle} to get its factor code prediction $\hat{c}'^{(n)} = E_{\text{oracle}}(x'^{(n)})$. Accordingly, we measure the correlation between $\hat{c}'^{(n)}$ and $c'^{(n)}$ in the same way with prior disentanglement metrics. In particular, we define an MIG-like metric, called *MIG-gen*, to evaluate the generator,

$$\text{MIG-gen} = \frac{1}{NK} \sum_{n=0}^{N-1} \sum_{k=0}^{K-1} \frac{1}{H(\hat{c}_k^{(n)})} \cdot \left(I(\hat{c}'^{(n)}_{j_k}; c'^{(n)}_k) - \max_{j \neq j_k} I(\hat{c}'^{(n)}_j; c'^{(n)}_k) \right)$$



(a) Isaac3D (b) Falcor3D

Figure 2. Semi-StyleGAN with the default setting $\gamma_G = \beta = \gamma$, $\gamma_E = 0$, $\alpha = 1$ where $\gamma \in \{1, 10\}$ on (a) Isaac3D and (b) Falcor3D. We vary the portion of labeled data η to show the impact of semi-supervision by comparing with Info-StyleGAN (i.e. the unsupervised baseline), and the fully-supervised one ($\eta = 1$). Only using 0.25∼2.5% of labeled data achieves near fully-supervised disentanglement.

Methods	MIG ↑	L2 ↓	MIG-gen ↑	L2-gen ↓
Encoder-only	0.731 ± 0.009	0.379 ± 0.002	-	-
Encoder-only w/ MixUp	0.834 ± 0.004	0.279 ± 0.005	-	-
Semi-StyleGAN	0.812 ± 0.020	0.301 ± 0.012	**0.965 ± 0.014**	**0.052 ± 0.016**
+ Remove smoothness consistency	0.765 ± 0.042	0.347 ± 0.019	0.945 ± 0.011	0.072 ± 0.008
+ Increase the $\mathcal{L}_{\text{unsup}}$ term in E ($\gamma_E = 10$)	**0.880 ± 0.120**	**0.225 ± 0.222**	0.888 ± 0.087	0.283 ± 0.247
+ Remove the $\mathcal{L}_{\text{unsup}}$ term in G	0.719 ± 0.014	0.490 ± 0.024	0.130 ± 0.054	1.514 ± 0.003

(a) Isaac3D (η = 0.1%)

Methods	MIG ↑	L2 ↓	MIG-gen ↑	L2-gen ↓
Encoder-only	0.690 ± 0.007	0.271 ± 0.002	-	-
Encoder-only w/ MixUp	0.701 ± 0.005	**0.265 ± 0.003**	-	-
Semi-StyleGAN	**0.704 ± 0.007**	0.285 ± 0.002	**0.754 ± 0.017**	**0.205 ± 0.022**
+ Remove smoothness consistency	0.674 ± 0.011	0.296 ± 0.017	0.632 ± 0.058	0.303 ± 0.088
+ Increase the $\mathcal{L}_{\text{unsup}}$ term in E ($\gamma_E = 10$)	0.643 ± 0.035	0.343 ± 0.016	0.636 ± 0.065	0.346 ± 0.070
+ Remove the $\mathcal{L}_{\text{unsup}}$ term in G	0.680 ± 0.016	0.300 ± 0.010	0.034 ± 0.028	1.096 ± 0.086

(b) Falcor3D (η = 0.5%)

Table 3. Ablation studies of Semi-StyleGAN on (a) Isaac3D and (b) Falcor3D, where the default setting is $\gamma_G = \beta = 10$, $\gamma_E = 0$, $\alpha = 1$. "Encoder-only" means we train the encoder by minimizing the L2 score with the labeled data only, a supervised baseline for the encoder disentanglement. "Encoder-only w/ MixUp" means we train the encoder by using MixUp (Zhang et al., 2018), a semi-supervised baseline for the encoder disentanglement. We set $\eta = 0.1\%$ on Isaac3D and $\eta = 0.5\%$ on Falcor3D, respectively.

where K is the length of factor code, $H(\cdot)$ and $I(\cdot;\cdot)$ denote the entropy and mutual information, respectively, and $j_k = \arg\max_j I(\hat{c}_j'^{(n)}, c_k'^{(n)})$. Similarly, we also introduce *L2-gen* to measure the semantic correctness of the generator,

$$\text{L2-gen} = \frac{1}{N}\sum_{n=0}^{N-1} \|E_{\text{oracle}}(x'^{(n)}) - c'^{(n)}\|_2$$

Intuitively, if the oracle encoder is perfect for every ground-truth observation-code pair, any mismatch between its prediction and the corresponding factor code should be contributed to the generator instead. Thus, both MIG-gen and L2-gen can effectively measure the generator controllability. To obtain an oracle encoder for each dataset, such as the proposed Falcor3D and Isaac3D, we pre-train a separate encoder network by minimizing the $L2$ score with all the ground-truth observation-code pairs.

4.3. Experimental Setup

Datasets and evaluation metrics. To test the proposed Semi-StyleGAN on complex high-resolution images that many prior works have difficulty with, we focus on three datasets: Isaac3D with resolution 512x512, Falcor3D with resolution 512x512 and CelebA with resolution 256x256. For the proposed Isaac3D and Falcor3D, we use *MIG* and *L2* to measure the encoder disentanglement, and *MIG-gen* and *L2-gen* to measure the generator controllability. For experiments on CelebA, we focus on the latent traversals to qualitatively measure the disentanglement quality.

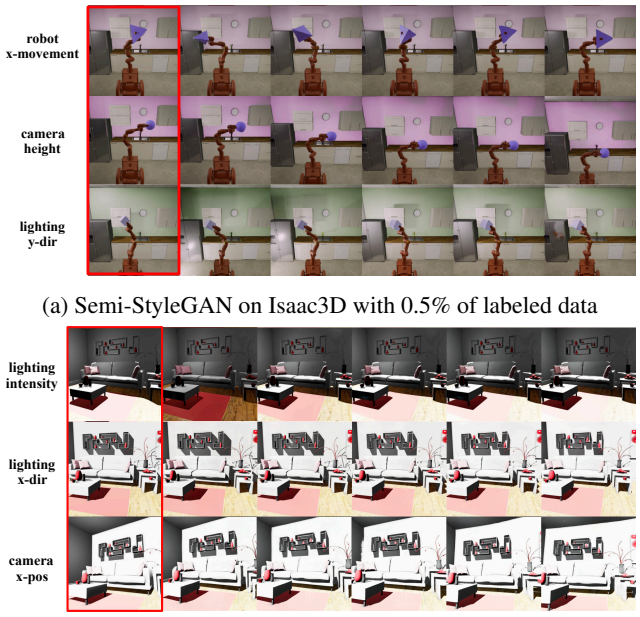
Experimental protocol. Before training, we first get the labeled set $\mathcal{J}$, by randomly sampling observation-code pairs from each dataset with a probability η. All the remaining observations form as the unlabeled set. The value of $\eta \in [0, 1]$ controls the portion of labeled data during training. Particularly Semi-StyleGAN becomes a fully-supervised method if $\eta = 1$, and reduces to Info-StyleGAN if $\eta = 0$. In experiments, we set $\xi = 0.75$ in Eq. (5) to be the same with (Berthelot et al., 2019). For the hyperparameters $\{\gamma_G, \gamma_E, \beta, \alpha\}$, we find that setting $\gamma_G = \beta = \gamma$, $\gamma_E = 0$, $\alpha = 1$ works well across different datasets, where we vary $\gamma \in \{1, 10\}$. Thus, without stated otherwise, we use the above setting by default in Semi-StyleGAN.

Our experiments mainly include four aspects. (i) We first vary the supervision rate η to show the impact of limited supervision. (ii) Given the supervision rate η, we train the encoder alone with the labeled data only (called *Encoder-only*) and with the MixUp (called *Encoder-only w/ MixUp*), respectively, as supervised and semi-supervised baselines for the encoder disentanglement. (iii) For ablation studies, we vary γ_G, γ_E to reveal the trade-off between the encoder and generator disentanglement, and vary α to show the impact of smoothness regularization. (iv) We show latent traversal results on both synthetic and real datasets.

4.4. Key Results

Impact of limited supervision. Figure 2 shows the quantitative results of varying η in Semi-StyleGAN on Isaac3D and Falcor3D, where we consider two cases of $\gamma \in \{1, 10\}$. For Isaac3D, only using 0.25% of the labeled data can achieve a very close performance to the fully-supervised one ($\eta = 1$) in terms of both the encoder and generator disentanglement. Similarly for Falcor3D, only using 2.5% of the labeled data can also achieve near the fully-supervised disentanglement. It means adding a very small amount of labeled data (0.25%$\sim$2.5%) into the training dataset could benefit significantly the disentanglement learning with Semi-StyleGAN. Besides, we can see that the generator disentanglement is more sensitive to the choice of γ than the encoder disentanglement, particularly on Falcor3D.

Ablation studies and comparison with baselines. First, Table 3 shows there may exist a crucial trade-off between the encoder and generator disentanglement, governed by the interplay between the unsupervised and supervised loss terms. For example, we can see that Semi-StyleGAN gets the best generator disentanglement by slightly sacrificing the encoder disentanglement, in particular on Isaac3D. Second, by removing the smoothness consistency, we can see a large performance drop in terms of all metrics on both datasets, which demonstrates the effectiveness of the smoothness regularization in Semi-StyleGAN. Besides, there exist a trade-off between the generator controllability and encoder disentanglement. For example, increasing the $\mathcal{L}_{\text{unsup}}$ term in E ($\gamma_E = 10$) worsens the generator controllability while improving the encoder disentanglement on Isaac3D. Removing the $\mathcal{L}_{\text{unsup}}$ term in G ($\gamma_G = 0$), we can still get a decent encoder disentanglement despite the generator controllability totally fails. This trade-off also depends on the datasets, evidenced by different behaviors on Isaac3D and Falcor3D. Finally, Table 3 shows the results of comparing Semi-StyleGAN with the supervised and semi-supervised baselines. We can see that Semi-StyleGAN significantly outperforms the supervised baseline (i.e., Encoder-only) on Isaac3D, and also gets a better MIG score on Falcor3D. With good hyperparameters which potentially weigh more on the encoder side, we can also achieve on par or better encoder disentanglement than the semi-supervised baseline (i.e., Encoder-only w/ MixUp).



(a) Semi-StyleGAN on Isaac3D with 0.5% of labeled data

(b) Semi-StyleGAN on Falcor3D with 1% of labeled data

Figure 3. Latent traversal of Semi-StyleGAN on Isaac3D and Falcor3D where $\gamma = 10$. Images in the first column (marked by red box) are randomly sampled real images and the rest images in each row are their interpolations, by uniformly varying the given factor from 0 to 1. See Appendix C.1 and C.2 for more results.

Latent traversal on synthetic and real data. Qualitatively, we show the latent traversal results on both synthetic and real data in Figure 3 and 4. When only 0.5% or 1% of the labeled data is available, each factor in the interpolated images changes smoothly without affecting other factors. For Isaac3D and Falcor3D, all the interpolated images visually look the same with their reference real image except for the considered factor, verifying the semantic correctness of Semi-StyleGAN with very limited supervision. For CelebA, we use a higher resolution than the prior work (Chen et al., 2016; Higgins et al., 2017; Nguyen-Phuoc et al.,

Figure 4. Latent traversal of Semi-StyleGAN on CelebA with resolution 256x256 by using 0.5% of the labeled data, where we use $\gamma = 1$ and disentangle all 40 binary attributes. See Appendix C.3 for the results of other attributes.

2019), and achieve visually better disentanglement quality together with a higher image quality. It means that the insights gained in the synthetic datasets also applies to the real domain. With very limited supervision, Semi-StyleGAN can achieve good disentanglement on real data.

5. An Extension for Better Generalization

Although Semi-StyleGAN performs well in both synthetic and real data, it cannot generalize to unseen data whose high-level content does not match the training data but whose fine-grained styles might. For instance, Semi-StyleGAN trained on Isaac3D cannot generate an image in which the robot arm stands on the right (instead of in the middle as in the training data). In this section, we design a new GAN architecture that extends Semi-StyleGAN to an image-to-image model, that we call Semi-StyleGAN-*fine*. This model achieves better generalization to unseen data.

Inspired by the observations that lower resolution blocks in the StyleGAN generator learn coarse-grained features while its high-resolution blocks account for fine-grained styles, we change the StyleGAN generator to not contain lower-resolution blocks. As shown in Figure 5, the generator instead takes the real image as one of its inputs by downscaling it to a lower resolution ϕ (e.g., $\phi = 32$ in Figure 5). Accordingly, it generates the high-resolution image by only modulating the (fine-grained) factor code into higher resolution blocks. Also, the encoder predicts the value of factor code from the block with resolution ϕ, instead of the last output block. The intuition is that lower-resolution blocks in the encoder also have less relationship with fine-grained styles, and thus the code prediction better use its higher resolution blocks only. This way, the generator in Semi-StyleGAN-*fine* does a semi-supervised controllable

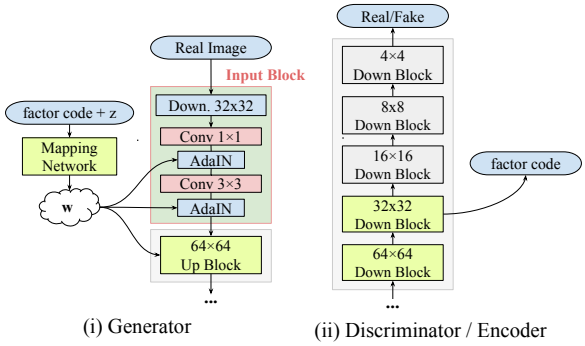


Figure 5. An illustration of Semi-StyleGAN-*fine*, where we downsample the real image into 32x32 resolution and replace the lower resolution blocks (4x4 - 32x32) in the generator by a new input block. Also, the encoder predicts the value of (fine-grained) factor code from the 32x32 block instead.

Figure 6. Generalized latent traversal results of Semi-StyleGAN-*fine* trained on Isaac3D with 1% of labeled data where we set $\phi = 64$ and interpolate the shown fine styles. In the test images, we shift the position of the robot arm to the right side, and also attach it with an unseen object (i.e., a tetrahedron).

fine-grained image editing while its encoder infers the fine-grained factor code that the generator has used. Finally, the loss functions remain the same as in Eq. (7).

5.1. Experimental Setup

We mainly focus on the latent traversals of Semi-StyleGAN-*fine* on Isaac3D and CelebA to qualitatively test its generalization ability. In the training time, we train Semi-StyleGAN-*fine* on Isaac3D and CelebA, respectively. In the test time, we apply novel test images (with the different high-level content) as the input to evaluate the proposed method. For Isaac3D, the novel test images are given by: (i) shifting the robot position to the right side (instead of standing right in the middle for all the training data), and (ii) attaching the robot arm with an unseen object. For CelebA, we simply download some new face images from online, following by aligning and cropping them into 256x256.

5.2. Key Results

Figure 6 and 7 show the results of interpolating fine-grained factors in the Isaac3D dataset and the CelebA dataset, respectively, with different test images where we set $\eta = 0.01$

Figure 7. Generalized latent traversal results of Semi-StyleGAN-*fine* trained on CelebA with 1% of labeled data where we set $\phi = 64$ and control the shown fine styles.

and $\phi = 64$. We can see that each considered fine-grained factor in both datasets keeps changing smoothly during its interpolations without affecting other factors, implying good generalized disentanglement. Particularly, the interpolations of the Isaac3D test images all maintain the new robot position and new object shape (i.e., a tetrahedron) with relatively high image quality. See Appendix D for test images with another novel object shape. The interpolations of the CelebA test images also keep the same identities and other coarse-grained features with the given input images. It is worthy to note that the good generalized disentanglement of Semi-StyleGAN-*fine* has been achieved by using only 1% of labeled data, and further increasing the supervision rate η does not visually improve performance. Therefore, these results demonstrate the ability of Semi-StyleGAN-*fine* in semantic fine-grained image editing with limited supervision that generalizes well to unseen novel images.

6. Conclusions

In this paper, we designed new loss functions and architectures based on StyleGAN for semi-supervised high-resolution disentanglement learning. We first showed that Info-StyleGAN largely outperforms most state-of-the-art unsupervised disentanglement methods, which justified the advantages of the StyleGAN architecture in disentanglement learning. We then proposed Semi-StyleGAN that achieved near fully-supervised disentanglement with limited supervision (0.25%$\sim$2.5%) on complex high-resolution synthetic and real data. We also proposed new metrics to quantify the generator controllability. To the best of our knowledge, we were the first to reveal that there exists a trade-off between learning disentangled representations and controllable generation. Besides, we extended Semi-StyleGAN to do semantic fine-grained image editing with better generalization to unseen images. Finally, we created two high-quality synthetic datasets to serve as new disentanglement benchmarks.

In the future, we want to apply Semi-StyleGAN to even larger-scale high-resolution real datasets. We are aware of the gender and racial biases in the CelebA dataset (Kärkkäinen & Joo, 2019), and thus hope to create better datasets and find other ways to address the algorithmic bias. Besides, It would be interesting to extend Semi-StyleGAN to the weakly-supervised, semi-supervised scenario, where the factors of variation are only partially observed.

Acknowledgement

Thanks to the anonymous reviewers for useful comments. We also thank Zhiding Yu, Anuj Pahuja, Yaosheng Fu, Tan Minh Nguyen and many others at Nvidia for helpful discussions on this work. WN and ABP were supported by IARPA via DoI/IBC contract D16PC00003.

References

Achille, A. and Soatto, S. Emergence of invariance and disentanglement in deep representations. *The Journal of Machine Learning Research*, 19(1):1947–1980, 2018.

Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., and Raffel, C. A. Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems*, 2019.

Bouchacourt, D., Tomioka, R., and Nowozin, S. Multi-level variational autoencoder: Learning disentangled representations from grouped observations. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Chen, T. Q., Li, X., Grosse, R. B., and Duvenaud, D. K. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems*, 2018.

Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., and Abbeel, P. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems*, 2016.

Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., and Choo, J. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018.

Gondal, M. W., Wüthrich, M., Miladinović, D., Locatello, F., Breidt, M., Volchkov, V., Akpo, J., Bachem, O., Schölkopf, B., and Bauer, S. On the transfer of inductive bias from simulation to the real world: a new disentanglement dataset. *arXiv preprint arXiv:1906.03292*, 2019.

He, Z., Zuo, W., Kan, M., Shan, S., and Chen, X. Attgan: Facial attribute editing by only changing what you want.

IEEE Transactions on Image Processing, 28(11):5464–5478, 2019.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, 2017.

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

Huang, X. and Belongie, S. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

Hyvärinen, A. and Pajunen, P. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429–439, 1999.

Kärkkäinen, K. and Joo, J. Fairface: Face attribute dataset for balanced race, gender, and age. *arXiv preprint arXiv:1908.04913*, 2019.

Karras, T., Aila, T., Laine, S., and Lehtinen, J. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

Kim, H. and Mnih, A. Disentangling by factorising. In *International Conference on Machine Learning*, 2018.

Kingma, D. P., Mohamed, S., Rezende, D. J., and Welling, M. Semi-supervised learning with deep generative models. In *Advances in neural information processing systems*, 2014.

Kulkarni, T. D., Whitney, W. F., Kohli, P., and Tenenbaum, J. Deep convolutional inverse graphics network. In *Advances in neural information processing systems*, 2015.

Kumar, A., Sattigeri, P., and Balakrishnan, A. Variational inference of disentangled latent concepts from unlabeled observations. *arXiv preprint arXiv:1711.00848*, 2017.

Lin, Z., Thekumparampil, K. K., Fanti, G., and Oh, S. Infogan-cr: Disentangling generative adversarial networks with contrastive regularizers. *arXiv preprint arXiv:1906.06034*, 2019.

Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schölkopf, B., and Bachem, O. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning*, 2019.

Locatello, F., Tschannen, M., Bauer, S., Rätsch, G., Schölkopf, B., and Bachem, O. Disentangling factors of variation using few labels. In *ICLR*, 2020.

Matthey, L., Higgins, I., Hassabis, D., and Lerchner, A. dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017.

Mirza, M. and Osindero, S. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

Miyato, T. and Koyama, M. cgans with projection discriminator. In *ICLR*, 2018.

Narayanaswamy, S., Paige, T. B., Van de Meent, J.-W., Desmaison, A., Goodman, N., Kohli, P., Wood, F., and Torr, P. Learning disentangled representations with semi-supervised deep generative models. In *Advances in Neural Information Processing Systems*, 2017.

Nguyen-Phuoc, T., Li, C., Theis, L., Richardt, C., and Yang, Y.-L. Hologan: Unsupervised learning of 3d representations from natural images. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

Odena, A., Olah, C., and Shlens, J. Conditional image synthesis with auxiliary classifier gans. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

Ridgeway, K. and Mozer, M. C. Learning deep disentangled embeddings with the f-statistic loss. In *Advances in Neural Information Processing Systems*, 2018.

Sajjadi, M., Javanmardi, M., and Tasdizen, T. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. In *Advances in neural information processing systems*, 2016.

Spurr, A., Aksan, E., and Hilliges, O. Guiding infogan with semi-supervision. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2017.

Ulyanov, D., Vedaldi, A., and Lempitsky, V. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

Xiao, T., Hong, J., and Ma, J. Dna-gan: Learning disentangled representations from multi-attribute images. *arXiv preprint arXiv:1711.05415*, 2017.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. Mixup: Beyond empirical risk minimization. In *ICLR*, 2018.