# InfoGAN-CR and ModelCentrality: Self-supervised Model Training and Selection for Disentangling GANs

**Zinan Lin** [1] **Kiran K. Thekumparampil** [2] **Giulia Fanti** [1] **Sewoong Oh** [3]

## Abstract

Disentangled generative models map a latent code vector to a target space, while enforcing that a subset of the learned latent codes are interpretable and associated with distinct properties of the target distribution. Recent advances have been dominated by Variational AutoEncoder (VAE)-based methods, while training disentangled generative adversarial networks (GANs) remains challenging. In this work, we show that the dominant challenges facing disentangled GANs can be mitigated through the use of self-supervision. We make two main contributions: first, we design a novel approach for training disentangled GANs with self-supervision. We propose *contrastive regularizer*, which is inspired by a natural notion of disentanglement: latent traversal. This achieves higher disentanglement scores than state-of-the-art VAE- and GAN-based approaches. Second, we propose an unsupervised model selection scheme called ModelCentrality, which uses generated synthetic samples to compute the medoid (multi-dimensional generalization of median) of a collection of models. The current common practice of hyper-parameter tuning requires using ground-truths samples, each labelled with known perfect disentangled latent codes. As real datasets are not equipped with such labels, we propose an unsupervised model selection scheme and show that it finds a model close to the best one, for both VAEs and GANs. Combining contrastive regularization with ModelCentrality, we improve upon the state-of-the-art disentanglement scores significantly, without accessing the supervised data.

[1]Carnegie Mellon University [2]University of Illinois at Urbana-Champaign [3]University of Washington. Correspondence to: Zinan Lin <zinanl@andrew.cmu.edu>, Kiran K. Thekumparampil <thekump2@illinois.edu>, Giulia Fanti <gfanti@andrew.cmu.edu>, Sewoong Oh <sewoong@cs.washington.edu>.

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

## 1. Introduction

The ability to learn low-dimensional, informative data representations can greatly enhance the utility of data. The notion of *disentangled* representations in particular was theoretically proposed in (Bengio et al., 2013; Ridgeway, 2016; Higgins et al., 2016) for diverse applications including supervised and reinforcement learning. A disentangled generative model takes a number of latent factors as inputs, with each factor controlling an interpretable aspect of the generated data. For example, in facial images, disentangled latent factors might control variations in eyes, noses, and hair.

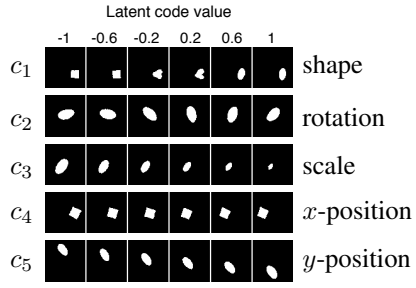


Figure 1: Each row shows how the image changes when traversing a single latent code under the proposed InfoGAN-CR architecture (dSprites dataset, § 3.2). Latent codes capture desired properties: {shape, rotation, scale, x-pos, y-pos}, of the image.

Most approaches for disentangling latent factors (or *codes*) are based on the following natural intuition. We say a generative model has a better disentanglement if changing one latent code (while fixing other latent codes) makes a *noticeable* and *distinct* change in the generated sample (referred to as "informativeness" and "disentanglement" in (Eastwood & Williams, 2018)). Noticeable changes are desired as we want the latent codes to capture important characteristics of the image. Distinct changes are desired as we want each latent code to represent an aspect of the samples different from other latent codes. (Eastwood & Williams, 2018) also values "completeness", which refers to how much of the disentangling latent factors are covered by the learned model. As such, disentanglement can be evaluated by traversing the

latent space as in Figure 1: by fixing all latent codes except one, varying that code, and visualizing the resulting changes. Figure 1 illustrates how the latent codes $\{c_1, \ldots, c_5\}$ of a successfully-trained generator capture noticeable and distinct properties of the images.

Two main obstacles arise in the design of disentangled generative models: (1) designing architectures that achieve good disentanglement *and* good sample quality, and (2) hyperparameter tuning and model selection given a fixed learning architecture.

For the first problem, recent approaches to disentanglement have focused on adding carefully chosen regularizers to promote disentanglement, building upon the two popular deep generative models: Variational AutoEncoders (VAE) (Kingma & Welling, 2013) and Generative Adversarial Networks (GAN) (Goodfellow et al., 2014). Fundamental differences in these two architectures led to the design of different regularizers. To achieve disentanglement in VAEs, a popular approach is to promote "*uncorrelatedness*" by regularizing with total correlation, as in $\beta$-VAE and FactorVAE (Higgins et al., 2016; Kim & Mnih, 2018). This approach has led to successful disentanglement scores, albeit at the cost of sample quality. Disentangled GANs, on the other hand, add a secondary input of latent codes, which are meant to control the underlying factors. The loss function then adds an extra regularizer to promote "*informativeness*", as proposed in InfoGAN (Chen et al., 2016). Despite improving sample quality, InfoGAN has lower disentanglement scores than its VAE-based counterparts, which led to slow progress on GAN-based disentangled representation learning.

The second problem, model selection, has received relatively less attention. Most prior work on disentanglement conducts hyperparameter tuning by cross-validating on a holdout dataset labelled with ground truth latent codes. This significantly limits the validity of those training methods on real datasets with unknown labels. However, recent work has acknowledged the need for unsupervised model selection techniques and proposed an unsupervised approach (Duan et al., 2019a). This approach was evaluated only on VAEs, and as we will show, it has poor performance on GAN-based models.

In summary, the two principal challenges associated with the design of disentangled generative models are particularly pronounced for disentangled GANs. This has contributed to a perception in the community that GANs are less well-suited to learning disentangled representations.

**Main contributions.** In this paper, we show that self-supervision can mitigate both of these challenges for disentangled GANs, allowing their performance to far supersede state-of-the-art VAE-based methods. We make two primary contributions:

First, we design a novel architecture for training disentangled GANs, which we call InfoGAN-CR. InfoGAN-CR adds a *contrastive regularizer* (CR) that combines self-supervision with the most natural measure of disentanglement: latent traversal. We create a self-supervised learning task of multi-way hypothesis tests over the latent codes and encouraging the generator to succeed at those tasks. We provide experimental results showing that it achieves state-of-the-art disentanglement scores on benchmark tasks.

Second, we introduce a novel model selection scheme based on self-supervision, which we call *ModelCentrality*. This builds upon a premise that well-disentangled models are close together, with the closeness measured by a popular disentanglement metric from (Kim & Mnih, 2018). We verify this premise numerically and define ModelCentrality as the medoid (multi-dimensional generalization of the median) of a set of models, computed under this disentanglement metric. ModelCentrality assigns centrality scores to each trained model based on the self-supervised labels defined by the closeness to other models . We demonstrate on benchmark datasets that ModelCentrality can be used for selecting both disentanlged GANs and VAEs. Models trained with InfoGAN-CR and selected with ModelCentrality significantly outperform state-of-the-art baseline approaches, even those that use supervised hyper-parameter tuning.

**Related work.** Learning a disentangled representation was first demonstrated in the *semisupervised setting*, where additional annotated data is available. This consists of examples from desired isolated latent factor traversals (Karaletsos et al., 2015; Kulkarni et al., 2015; Narayanaswamy et al., 2017; Lopez et al., 2018; Watters et al., 2019; Locatello et al., 2019b; Chen & Batmanghelich, 2019). However, as manual data annotation is costly, *unsupervised methods* for disentangling are desired. Early approaches to unsupervised disentangling imposed uncorrelatedness by making it difficult to predict one representational unit from the rest (Schmidhuber, 1992), disentangling higher order moments (Desjardins et al., 2012), using factor analysis (Tang et al., 2013), and applying group representations (Cohen & Welling, 2014). Breakthroughs in making these ideas scalable were achieved by $\beta$-VAE (Higgins et al., 2016) for VAE-based methods, and InfoGAN (Chen et al., 2016) for GAN-based ones. Rapid progress in improving disentanglement was driven mainly by VAE-based methods, in a series of papers (Kim & Mnih, 2018; Locatello et al., 2018; Chen et al., 2018; Lopez et al., 2018; Ansari & Soh, 2018; Esmaeili et al., 2018; Gao et al., 2018; Pineau & Lelarge, 2018; Dupont, 2018; Ainsworth et al., 2018b;a; Szabó et al., 2017; Burgess et al., 2018; Jeong & Song, 2019; Li et al., 2019; Caselles-Dupré et al., 2019; Tschannen et al., 2018). Quantitative comparisons in these papers suggest that InfoGAN learns poorly-disentangled representations. This has led to a misconception that GAN-based methods are

inherently bad at learning disentangled representations.

Concurrent and subsequent to our work, several other GAN-based disentangling frameworks have been proposed (Jeon et al., 2018; Liu et al., 2019; Lee et al., 2020) and these work corroborate our finding that VAE-based approaches are not superior in disentangling. Additionaly, various domain specific models have also been proposed for structured data such as sequences (Hsu et al., 2017), images (Awiszus et al., 2019; Lee et al., 2018), video (Xing et al., 2018; Denton et al., 2017; Hsieh et al., 2018), shapes (Aumentado-Armstrong et al., 2019; Lorenz et al., 2019), and state space (Miladinović et al., 2019). Several works have studied the use of disentangled representations in diverse topics such as transfer learning (Higgins et al., 2017), hierarchical visual concepts learning (Higgins et al., 2018b), visual reasoning (van Steenkiste et al., 2019), fairness of learning (Locatello et al., 2019a; Marx et al., 2019; Creager et al., 2019), computer vision (Lee et al., 2018; Hsieh et al., 2018; Singh et al., 2019), speech processing (Hsu et al., 2017), robust learning (Duan et al., 2019b).

## 2. Background

In this section, we give a brief overview of GANs and InfoGAN, introduced in (Chen et al., 2016).

**Background on GAN.** Generative Adversarial Networks (GANs) are a breakthrough method for training generative models (Goodfellow et al., 2014). A deep neural network generative model maps a latent code $z \in \mathbb{R}^d$ to a desired distribution of the samples $x = G(z)$. $z$ is typically drawn from a Gaussian distribution with identity covariance or a uniform distribution. No likelihood is available for ML training of the neural network $G$. GANs instead update weights of a generator $G$ and discriminator $D$ using alternative gradient updates on the following *adversarial loss*:

$$\min_G \max_D \ \mathcal{L}_{\mathrm{Adv}}(D, G) \ . \tag{1}$$

The discriminator provides an approximate measure of how different the current generator distribution is from the distribution of the real data. For example, a common choice is $\mathcal{L}_{\mathrm{Adv}}(D, G) = \mathbb{E}_{x \sim P_{\mathrm{real}}}[\log(D(x))] + \mathbb{E}_{P_G}[\log(1 - D(x))]$, which provides an approximation of the Jensen-Shannon divergence between the real data distribution $P_{\mathrm{data}}$ and the current generator distribution $P_G$.

**Background on InfoGAN.** In order to achieve disentanglement, InfoGAN proposes a regularizer based on mutual information. As the goal is not to disentangle all latent codes, but rather to disentangle a subset, InfoGAN (Chen et al., 2016) proposed to first split the latent codes into two parts: the disentangled code vector $c \in \mathbb{R}^k$ and the remaining code vector $z \in \mathbb{R}^d$ that provides additional randomness. InfoGAN then uses the GAN loss with regularization to encourage informative latent codes $c$:

$$\min_G \max_D \ \mathcal{L}_{\mathrm{Adv}}(G, D) - \lambda \, I(c; G(c, z)) \ , \tag{2}$$

where $I(c; G(c, z))$ denotes the mutual information between the latent code $c$ and the sample $G(c, z)$ generated from that latent code, and $\lambda$ is a positive scalar coefficient. Notice that encouraging informativeness alone does not necessarily imply good disentanglement; a fully entangled representation can achieve infinite mutual information $I(c; G(c, z))$. Despite this, InfoGAN achieves reasonable performance in practice. Its decent empirical performance follows from implementation choices that promote stability and alter the InfoGAN objective, which we discuss in Appendix A.

## 3. Self-supervision with Contrastive Regularizer

Our proposed regularizer is inspired by the idea that disentanglement should be measured via changes in the images when traversing the latent space. This is a popular interpretation of disentanglement, as evidenced by the widely-adopted visual evaluations (e.g. Figure 1). This suggests a natural disentanglement approach: run latent traversal experiments and encourage models that make *distinct* changes.

We design a regularizer, which we call a *Contrastive Regularizer* (CR), based on this insight. That is, we generate two (or more) images from the generator, while fixing one of the latent codes $c_i$ to be the same for both images. We draw the rest of the latent codes uniformly at random, and let $(x, x') \sim Q^{(i)}$ denote the resulting distribution of paired samples when factor $c_i$ is fixed. We propose measuring the distinctness of this latent traversal with Jensen-Shannon divergence among $Q^{(i)}$'s defined as

$$d_{\mathrm{JS}}(Q^{(1)}, \ldots, Q^{(k)}) \triangleq \frac{1}{k} \sum_{i \in [k]} d_{\mathrm{KL}}\big(Q^{(i)} \,\big|\big|\, \bar{Q}\big) \, , \tag{3}$$

where $\bar{Q} = (1/k) \sum_{j \in [k]} Q^{(j)}$. This measures how different each latent code traversal is. If we maximize this as a regularizer to the generator training, in the subsequent generator update, the $Q^{(i)}$'s will be forced to be as different as possible. This, in turn, forces the changes in the latent codes to make changes in the images that are noticeable and easy to distinguish from (the changes of) other latent codes.

In general different ways of *coupling* the latent space of the paired images can be used, and we leave it as a design choice. For example, one could fix the rest of the codes to be the same and randomly sample $c_i$. This fits a more traditional definition of latent traversal. We provide practical guidelines in §3.2.

Before explaining how to implement our contrastive regularizer in §3.1, we show in Figure 2 how it enhances

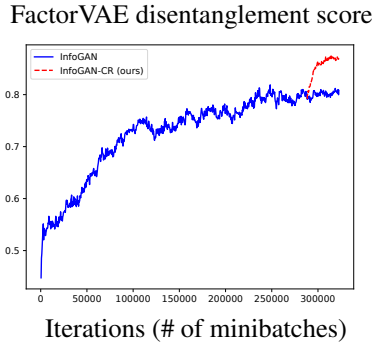


Figure 2: After 288,000 iterations, we continue training InfoGAN with(out) the proposed contrastive regularizer. The jump illustrates gains due to CR regularization. Curves are averaged over 10 trials on the same data.

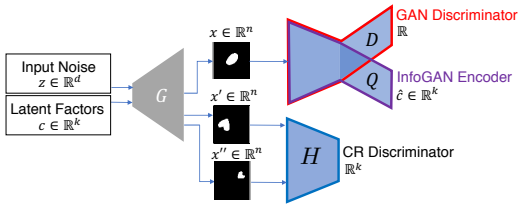


Figure 3: Like InfoGAN, InfoGAN-CR includes a GAN discriminator $D$ and an encoder $Q$, which share all convolutional layers and have separate fully-connected final layers. In addition, the CR discriminator $H$ takes as input a pair of images $x$ and $x'$ that are generated by sharing one fixed latent factor $c_i = c'_i$ for a randomly chosen $i \in [k]$, and randomly drawing the rest. The discriminator is trained to correctly identify $i$, the index of the fixed factor.

disentanglement beyond vanilla InfoGAN. The blue curve shows the performance when we train a vanilla InfoGAN on the dSprites dataset (Matthey et al., 2017) for 28 epochs (322,560 iterations) total. To show the effect of the proposed CR regularizer, we take the model we just trained with InfoGAN at 25 epochs (288,000 iterations), and keep training with an added CR-regularizer (red curve), precisely defined in Eq. (5). All other hyperparameters are identical. We measure disentanglement using the popular metric of (Kim & Mnih, 2018) and defined in §3.2. The jump at epoch 28 suggests that contrastive regularization significantly enhances disentanglement, on top of what was achieved by InfoGAN regularizer alone.

### 3.1. Contrastive Regularizer Architecture

To approximate the Contrastive Regularizer in (3), we introduce an additional discriminator $H : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^k$ that performs multi-way hypothesis testing. We then justify its use via an equivalence in an ideal scenario in Theorem 1. Building upon InfoGAN's architecture (see §2 for details), we add contrastive regularization and refer to the resulting architecture as InfoGAN-CR, illustrated in Figure 3. For non-negative scalars $\lambda$ and $\alpha$, this architecture is trained as

$$\min_{G,H,Q} \max_{D} \; \mathcal{L}_{\text{Adv}}(G, D) - \lambda \mathcal{L}_{\text{Info}}(G, Q) - \alpha \mathcal{L}_{\text{c}}(G, H) \quad (4)$$

The pair of coupled images $x$ and $x'$ are generated according to a choice of a coupling that defines how to traverse the latent space. The discriminator $H$ tries to identify which code $i$ was shared between the paired images. Both the generator and the discriminator try to make the $k$-way hypothesis testing successful. We use the standard cross entropy loss:

$$\mathcal{L}_{\text{c}}(G, H) = \mathbb{E}_{I \sim \text{U}([k]),(x,x') \sim Q^{(I)}} [\langle I \, , \, \log H(x, x') \rangle] \, , \quad (5)$$

where $Q^{(I)}$ denotes the joint distribution of the paired images, $I$ denotes the one-hot encoding of the random index, and $H$ is a $k$-dimensional vector-valued neural network normalized to be $\langle \mathbb{1}, H(x, x') \rangle = 1$ for all $x$ and $x'$. This naturally encourages each latent code to make distinct and noticeable changes, hence promoting disentanglement. Further, the following theorem justifies the use of this architecture and loss. We provide a proof in Appendix C.

**Theorem 1.** *When maximized over the class of all functions, the maximum of Eq.* (5) *is achieved by* $H(x, x') = (1/Z_{x,x'}) \left[ Q^{(1)}(x, x') \quad , \cdots , \quad Q^{(k)}(x, x') \right]$ *with a normalizing constant* $Z_{x,x'} = \sum_{i \in [k]} Q^{(i)}(x, x')$ *and the maximum value is the generalized Jensen-Shannon divergence,*

$$\max_{H} \mathcal{L}_c(G, H) = d_{\text{JS}}(Q^{(1)}, \ldots, Q^{(k)}) - \log k \, .$$

**Progressive training.** There are many ways to couple the latent variables. We prescribe progressively changing the hypotheses (or how we couple the images) during the course of the training, from easy to hard. The hypotheses class we propose is as follows. First we draw a random index $I$ over $k$ indices, and sample the chosen latent code $c_I \in \mathbb{R}$. Two images are generated with the same value of $c_I$; the remaining factors are chosen independently at random. Letting $c_j^m$ denote the $j$th latent code for image $m \in \{1, 2\}$, the *contrastive gap* is defined as $\min_{j \in [k] \setminus \{I\}} |c_j^1 - c_j^2|$. In Appendix D, we discuss in more detail how we sample the latent codes for a given choice of a contrastive gap. The larger the contrastive gap, the more distinct the pair of samples. We gradually reduce the contrastive gap for progressive training (§3.2.1). Figure 4 illustrates the power of progressive training on dSprites dataset. For the 'progressive training' curve, we use a contrastive gap of 1.9 for 120,000 batches, and then introduce a (more aggressive) gap of 0. For the 'no progressive training' curves, we use gap size of 0 or 1.9 for all 230,400 batches.

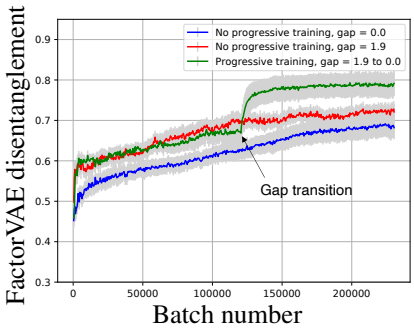


Figure 4: Reducing the contrastive gap from 1.9 to 0 during training significantly improves FactorVAE scores.

### 3.2. Empirical Evaluation of Contrastive Regularizer with Supervised Hyper-parameter Tuning

For quantitative evaluation, we run experiments on synthetic datasets with pre-defined latent factors, including dSprites (Matthey et al., 2017) and 3DTeapots (Eastwood & Williams, 2018).[1] We evaluate disentanglement using the popular metrics from (Kim & Mnih, 2018; Eastwood & Williams, 2018; Kumar et al., 2017; Ridgeway & Mozer, 2018; Chen et al., 2018; Higgins et al., 2016). For qualitative evaluation, we use our synthetic datasets as well as the CelebA dataset (Liu et al., 2015). More details on datasets and metrics can be found in Appendix E.

It is typical in disentanglement literature to select hyperparameters in a supervised manner in synthetic datasets where ground truth disentanglement is known. We do the same in this section and choose hyperparameters of all the models we train (FactorVAE, InfoGAN modified, and InfoGAN-CR). These are fair comparisons as all reported scores in this section are results of such hyperparameter tuning (some by us and some by the experimenters). However, this practice of supervised hyperparameter tuning is problematic; we resolve this issue in §4. Perhaps surprisingly, we show that our *unsupervised model selection* finds a better model than that found via supervised hyperparameter tuning.

#### 3.2.1. dSprites Dataset

We compute and/or reproduce disentanglement metrics for a number of protocols in Table 1. We provide details of the experiments in Appendix F, and focus on the interpretation of the results in this section. An example of latent traversal of the output of InfoGAN-CR is shown in Figure 1.

Contrastive regularization provides a clear gain in disentanglement, bringing InfoGAN-CR's FactorVAE score up to 0.90, higher than any baseline from the VAE or GAN literature. A similar trend holds for most of the metrics. We were made aware of independent work that proposes a special case of Contrastive Regularization in (Li et al., 2018); concretely, (Li et al., 2018) fixes $\lambda = 0$ in our loss (4), and also uses a special coupling that matches all but one latent code in $c$ for the matched pairs. This empirically achieves a lower FactorVAE scores ($0.39 \pm 0.02$ standard error over 10 runs) than even vanilla InfoGAN. Note that this difference is not a matter of parameter tuning, but of the loss function and training mechanism; indeed, in our own preliminary trials, we found that training a CR-regularizer without the InfoGAN loss, as in (Li et al., 2018), achieved similarly poor performance. The choice of coupling in our contrastive regularizer, the progressive training we propose, and the InfoGAN loss are all critical in achieving the improved the performance, as described in Appendix F.5. Hence, we do not consider it as a baseline moving forward.

#### 3.2.2. 3DTeapots Dataset

We ran InfoGAN-CR on the 3DTeapots dataset from (Eastwood & Williams, 2018), with images of teapots in various orientations and colors generated by the renderer in (Moreno et al., 2016). Details on this point, our implementation, and additional plots appear in Appendix G. Table 2 shows the disentanglement scores of FactorVAE and InfoGAN compared to InfoGAN-CR. While the results with this *supervised hyperparameter tuning* are mixed (none of the methods dominate), we show in §4, Table 4 that, perhaps surprisingly, our proposed *unsupervised model selection* finds a model that dominates all baseline algorithms.

#### 3.2.3. CelebA Dataset

We train InfoGAN-CR on the CelebA dataset of 202,599 celebrity facial images. Since these images do not have known continuous latent factors, we cannot compute the disentanglement metric. We therefore evaluate this dataset qualitatively by producing latent traversals, as seen in Figure 5. Details of this experiment are included in Appendix I.

## 4. ModelCentrality: Self-supervised Model Selection

The achievable scores in Table 1 are a consequence of *supervised* hyper-parameter tuning, for both our models and all baseline models. As shown in Figure 6, the designer runs experiments with multiple hyper-parameters—whose performance could vary significantly—and chooses one hyper-parameter that gives the best average performance. This approach is supervised, as performance evaluation requires access to a synthetic data generator with access to the ground truth disentangled codes.

Supervised hyper-parameter tuning is problematic, as $(i)$ in important real-world applications we do not have ground

[1] The code for all experiments is available at `https://github.com/fjxmlzn/InfoGAN-CR`

| | Model | FactorVAE | DCI | SAP | Explicitness | Modularity | MIG | BetaVAE |
|---|---|---|---|---|---|---|---|---|
| **VAE** | VAE | $0.63 \pm .06$ | $0.30 \pm .10$ | | | | 0.10 | |
| | $\beta$-TCVAE | $0.62 \pm .07$ | $0.29 \pm .10$ | | | | **0.45** | |
| | HFVAE | $0.63 \pm .08$ | $0.39 \pm .16$ | | | | | |
| | $\beta$-VAE | $0.63 \pm .10$ | $0.41 \pm .11$ | 0.55 | | | 0.21 | |
| | CHyVAE | 0.77 | | | | | | |
| | DIP-VAE | | | 0.53 | | | | |
| | FactorVAE | 0.82 | | | | | 0.15 | |
| | FactorVAE (1.0) | $0.79 \pm .01$ | $0.67 \pm .03$ | $0.47 \pm .03$ | $0.78 \pm .01$ | $0.79 \pm .01$ | $0.27 \pm .03$ | $0.79 \pm .02$ |
| | FactorVAE (10.0) | $0.83 \pm .01$ | $0.70 \pm .02$ | $0.57 \pm .00$ | $0.79 \pm .00$ | $0.79 \pm .00$ | $0.40 \pm .01$ | $0.83 \pm .01$ |
| | FactorVAE (20.0) | $0.83 \pm .01$ | $0.72 \pm .02$ | $0.57 \pm .00$ | $0.79 \pm .00$ | $0.79 \pm .01$ | $0.40 \pm .01$ | $0.85 \pm .00$ |
| | FactorVAE (40.0) | $0.82 \pm .01$ | $\mathbf{0.74 \pm .01}$ | $0.56 \pm .00$ | $0.79 \pm .00$ | $0.77 \pm .01$ | $0.43 \pm .01$ | $0.84 \pm .01$ |
| **GAN** | InfoGAN | $0.59 \pm .70$ | $0.41 \pm .05$ | | | | 0.05 | |
| | IB-GAN | $0.80 \pm .07$ | $0.67 \pm .07$ | | | | | |
| | InfoGAN (modified) | $0.82 \pm 0.01$ | $0.60 \pm 0.02$ | $0.41 \pm 0.02$ | $0.82 \pm 0.00$ | $0.94 \pm 0.01$ | $0.22 \pm 0.01$ | $0.87 \pm 0.01$ |
| | InfoGAN-CR | $\mathbf{0.88 \pm 0.01}$ | $0.71 \pm 0.01$ | $\mathbf{0.58 \pm 0.01}$ | $\mathbf{0.85 \pm 0.00}$ | $\mathbf{0.96 \pm 0.00}$ | $0.37 \pm 0.01$ | $\mathbf{0.95 \pm 0.01}$ |

Table 1: Comparisons of the popular disentanglement metrics on the dSprites dataset. A perfect disentanglement corresponds to 1.0 scores. The proposed InfoGAN-CR achieves the highest score on most cases, compared to the best reported result for each baseline. See Appendix A for InfoGAN (modified). The InfoGAN (modified) and InfoGAN-CR rows are averaged over 50 runs. Appendix F.2 gives more details on the reproducibility of the results. We show in Table 3 that with our proposed model selection scheme, we improve the performance even further.

| Model | FactorVAE | DCI | SAP | Explicitness | Modularity | MIG | BetaVAE |
|---|---|---|---|---|---|---|---|
| FactorVAE | $0.79 \pm .03$ | $0.55 \pm .04$ | $0.49 \pm .05$ | $\mathbf{0.84 \pm .01}$ | $0.72 \pm .02$ | $0.24 \pm .03$ | $\mathbf{0.94 \pm .02}$ |
| InfoGAN (modified) | $0.76 \pm .06$ | $0.62 \pm .06$ | $\mathbf{0.57 \pm .06}$ | $0.82 \pm .04$ | $\mathbf{0.98 \pm .01}$ | $0.34 \pm .04$ | $0.90 \pm .07$ |
| InfoGAN-CR | $\mathbf{0.82 \pm .02}$ | $\mathbf{0.66 \pm .01}$ | $0.53 \pm .02$ | $0.81 \pm .01$ | $0.97 \pm .00$ | $\mathbf{0.38 \pm .02}$ | $0.89 \pm .02$ |

Table 2: Comparisons of the popular disentanglement metrics on the 3DTeapots. We show in Table 4 that with our proposed model selection scheme, we achieve the best performance on all metrics.

truth data, and $(ii)$ a more complex model with a larger space to tune could get better scores by an extensive search. To this end, we propose a novel unsupervised model selection scheme called *ModelCentrality* that bypasses both of these concerns.

### 4.1. ModelCentrality

Suppose there is a notion of an optimal disentanglement that we want to discover from the data. We start from a premise that well-disentangled models should be close to that optimal model, and hence also close to each other. To measure similarity between models, we borrow insights from a long line of research in measuring disentanglement. In particular, prior work suggests that models with good disentanglement metrics (e.g. those in Table 1) tend to exhibit qualitatively good disentanglement properties, e.g., via latent traversals. This suggests that disentanglement scores can be used to measure how close the disentangled latent codes of one model are to the latent codes of another.

Consider a trained model $G : \mathbb{R}^k \to \mathbb{R}^n$ (here we only consider the model as mapping a disentangled latent code $c \in \mathbb{R}^k$ to the image $x \in \mathbb{R}^n$ and treat the $z \in \mathbb{R}^d$ as an inherent randomness in the generative model). Existing metrics also require the corresponding *encoder* $Q : \mathbb{R}^n \to \mathbb{R}^k$ that maps samples to estimated disentangled latent factors. For example, the popular FactorVAE metric of (Kim & Mnih, 2018) of a trained generative model $G_i$ measures how well its encoder $Q_i$ can estimate, from real samples, the true latents of real samples (for example the ground truths dSprites dataset with also the true disentangled latent factors).

Instead of the original FactorVAE score, which requires supervision from the training data with ground truths latent codes, we use other trained models as a surrogate for the ground truths. ModelCentrality treats the distribution of another model $G_j$ as the ground truth. Given two trained models: $G_i$ and $G_j$, we can measure how well the encoder $Q_i$ can estimate, from the generated samples of model $G_j$, the learned latents of the generated samples of model $G_j$. Hence, we can compute the similarity from $G_j$ to $G_i$ by (1) generating samples using the target model $G_j$; (2) passing those samples through the encoder $Q_i$ of model $G_i$ to estimate its latents, (3) using these estimated latents to evaluate the FactorVAE metric by using the latents generated by target model $G_j$ as ground truths. This similarity metric is an instance of self-supervision, as we treat one model as the *target label* and no ground-truth labels are needed.


Figure 5: Latent traversal for CelebA dataset, using InfoGAN-CR.

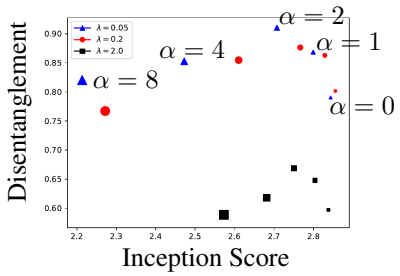


Figure 6: Inception score and FactorVAE score achieved by various hyper-parameters in (4). The size of each point denotes $\alpha \in \{0, 1, 2, 4, 8\}$, in the order of increasing size. We explicitly label this for $\lambda = 0.05$ (blue triangles).

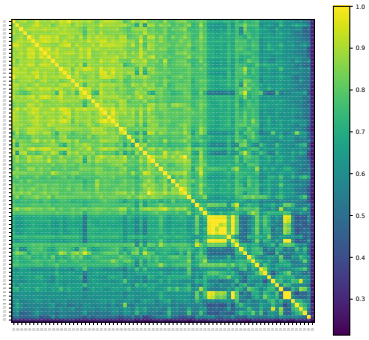

Figure 7: Heat map of the matrix $B$ used to compute ModelCentrality for each InfoGAN-CR model trained with dSprites dataset. Each row/column corresponds to one trained model, which are sorted according to FactorVAE score on the ground truth factors (computed with the supervised ground truths dSprites dataset). Top-left is the highest FactorVAE scoring model.

Given a pool of $N$ trained generative models, we compute $A_{ij}$ as the disentanglement score achieved by model $G_i$ treating model $G_j$ as the target model with the way mentioned above. Then we define a symmetric similarity matrix $B \in \mathbb{R}^{N \times N}$, where the similarity between a model $i$ and model $j$ is denoted by $B_{ij}$, and is computed as $B_{ij} = (1/2)(A_{ij} + A_{ji})$. In our experiments, we choose FactorVAE score as the disentanglement metric, because it is popular and robust, but we will show that FactorVAE-based ModelCentrality predicts all other scores accurately in Figure 8.

We experimentally confirm our premise that good models are close to each other in Figure 7, which illustrates the similarity matrix $B$. The rows/columns of this matrix are sorted by FactorVAE score, computed on the ground truth disentanglement factors. As expected, models that are better disentangled (as measured by FactorVAE score) are closer to each other (top-left), and the models that are not disentangled are far from other models (bottom-right).

This observation naturally suggests using some notion of a *central model* in our pool as the best model. We propose a measure of ModelCentrality based on the medoid of models with respect to the similarity matrix $B$. We define the ModelCentrality of a model $i$ as $s_i = \frac{1}{n-1} \sum_{j \neq i} B_{ij}$. We then select the model with the largest ModelCentrality, which coincides with the medoid in the pool of models. The pseudocode for computing ModelCentrality is given in Algorithm 1.

Besides model selection, ModelCentrality can be used for other tasks, as $s_i$ provides a quantitative evaluation of the $i$-th model. For example, ModelCentrality can rank the models according to $s_i$. It can also be used for hyper-parameter selection by averaging the $s_i$'s of the models trained with the same hyper-parameter, and selecting the best hyper-parameter.

### 4.2. Comparison with State-of-the-art Model Selection

We compare our model selection approach with state-of-the-art schemes from (Duan et al., 2019a). The first scheme,

**Algorithm 1** ModelCentrality

**Input:** $N$ pairs of generative models and latent code encoder: $(G_1, Q_1), ..., (G_N, Q_N)$,
supervised disentanglement metric $f$ : encoder $\times$ model $\to \mathbb{R}$
**Output:** the estimated best model $G^*$
Initialize a zero matrix: $A \in \mathbb{R}^{N \times N}$
**for** $i, j = 1 \to N$ **do**
  $A_{ij} \leftarrow f(Q_i, G_j)$
**end**
$B \leftarrow (A + A^T)/2$
**for** $i = 1 \to N$ **do**
  $s_i \leftarrow (\sum_{j \neq i} B_{ij})/(N-1)$
**end**
$k \leftarrow \arg\max_i s_i$
$G^* \leftarrow G_k$

UDR Lasso, defines a distance $A_{ij}$ from one model $i$ to another model $j$ as follows. Consider the encoder of model $i$ that maps an image to a latent code: $\hat{c} = Q_i(x) \in \mathbb{R}^k$. A linear regressor is trained with Lasso to predict $Q_i(x)$ from $Q_j(x)$ using samples $\{x^{(\ell)} \in \mathbb{R}^n\}_{\ell \in S_{\text{train}}}$ from the training dataset. If two models are identical, then the resulting (matrix valued) Lasso regressor will be a permutation matrix. Otherwise, a formula is applied to give a score (Duan et al., 2019a). The second approach, UDR Spearman, uses similar approach, except instead of training a Lasso regressor, the Spearman correlation coefficient is computed.

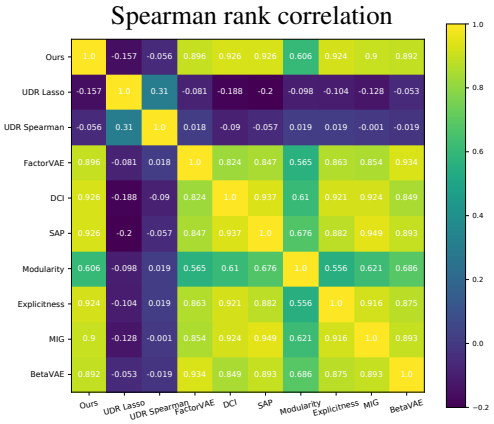


Figure 8: The rank correlation of the metrics on dSprites dataset. The first row/column is our ModelCentrality, which is highly correlated with all other disentanglement scores (row/column 4-10). Competing schemes of UDR Lasso and UDR Spearman are nearly uncorrelated.

In our experiments, we compare ModelCentrality to UDR Lasso and UDR Spearman on InfoGAN-CR and FactorVAE models trained on dSprites and 3DTeapots datasets. On the dSprites dataset, we first generated $N = 76$ InfoGAN-CR models from a grid of hyper-parameters. Figure 8 shows the Spearman rank correlations between models selected different metrics (including two UDR approaches and ModelCentrality). To produce this figure, we start with trained models $m_1, \ldots, m_{76}$, and a list of disentanglement metrics $f_1, \ldots, f_{10}$, including ModelCentrality, UDR (Spearman and Lasso), and an assortment of other disentanglement metrics. Then for the $i$th metric $f_i$, we compute $v_i = [f_i(m_1), \ldots, f_i(m_{76})] \in \mathbb{R}^{76}$. Note that all the metrics, except ModelCentrality and UDR, require the access to ground truths latent factors (and hence are supervised). Finally, the $(i, j)$th entry of Figure 8 is the Spearman rank correlation coefficient between vectors $v_i$ and $v_j$.

Figure 8 illustrates two points. First, ModelCentrality is not closely correlated with UDR Spearman or UDR Lasso, since the 2nd and 3rd rows/columns have low correlation coefficients. Second, ModelCentrality *is* closely correlated with the remaining disentanglement metrics. In Appendix K, we show a more detailed statistics of the scores, and show that a similar results hold when selecting FactorVAE models and also under 3DTeapots dataset. This suggests that choosing a model with maximum ModelCentrality tends to maximize existing disentanglement metrics, without requiring access to ground truth labels—an intuition that we confirm in Tables 3 and 4. Perhaps surprisingly, not only does ModelCentrality outperform UDR schemes, but it also selects models that outperform (a set of) models trained with a supervised hyper-parameter tuning from literature and from our experiments in §3.2. Notice the subtle difference in ModelCentrality producing a single model versus supervised hyper-parameter tuning producing a hyper-parameter for training a set of models. In fact, as shown in Table 3 and Table 4, the model selected with ModelCentrality has very close performance to the model with the best ground truth FactorVAE score. Under some metrics other than FactorVAE score, the model selected with ModelCentrality is even better.

A natural question is why ModelCentrality outperforms UDR. Several aspects of UDR Lasso contribute to its unreliability. $(i)$ Lasso involves a hyper-parameter, which can significantly change the resulting score. $(ii)$ Lasso is restricted to linear relations, whereas two perfectly disentangled models can have highly non-linear relations. $(iii)$ In addition, UDR Lasso does not generalize to discrete latent codes. UDR Spearman uses the Spearman's rank correlation in place of Lasso, and is reported to be inferior to UDR Lasso (Duan et al., 2019a). Notice that UDR schemes inherit the issues present in the disentanglement scores of DCI (Eastwood & Williams, 2018), from which the UDR schemes are derived. The proposed ModelCentrality is derived from FactorVAE scores (Kim & Mnih, 2018), which is popular, principled, and demonstrated to be a stable measure of disentanglement.

| Model | FactorVAE | DCI | SAP | Explicitness | Modularity | MIG | BetaVAE |
|---|---|---|---|---|---|---|---|
| FactorVAE with | | | | | | | |
| UDR Lasso | $0.81 \pm .00$ | $0.70 \pm .01$ | $0.56 \pm .00$ | $0.79 \pm .00$ | $0.78 \pm .00$ | $\mathbf{0.40 \pm .00}$ | $0.84 \pm .00$ |
| UDR Spearman | $0.79 \pm .00$ | $\mathbf{0.73 \pm .00}$ | $0.53 \pm .01$ | $0.79 \pm .00$ | $0.77 \pm .00$ | $\mathbf{0.40 \pm .01}$ | $0.79 \pm .00$ |
| ModelCentrality | $\mathbf{0.84 \pm .00}$ | $\mathbf{0.73 \pm .01}$ | $\mathbf{0.58 \pm .00}$ | $\mathbf{0.80 \pm .00}$ | $\mathbf{0.82 \pm .00}$ | $0.37 \pm .00$ | $\mathbf{0.86 \pm .00}$ |
| Best model in the pool | 0.88 | $nan$ | 0.58 | 0.79 | 0.79 | 0.39 | 0.83 |
| InfoGAN-CR with | | | | | | | |
| UDR Lasso | $0.86 \pm .01$ | $0.68 \pm .01$ | $0.49 \pm .01$ | $0.84 \pm .00$ | $0.96 \pm .00$ | $0.30 \pm .01$ | $0.92 \pm .01$ |
| UDR Spearman | $0.84 \pm .01$ | $0.67 \pm .01$ | $0.53 \pm .01$ | $0.84 \pm .00$ | $0.96 \pm .00$ | $0.31 \pm .01$ | $0.90 \pm .01$ |
| ModelCentrality | $\mathbf{0.92 \pm .00}$ | $\mathbf{0.77 \pm .00}$ | $\mathbf{0.65 \pm .00}$ | $\mathbf{0.87 \pm .00}$ | $\mathbf{0.99 \pm .00}$ | $\mathbf{0.45 \pm .00}$ | $\mathbf{0.99 \pm .00}$ |
| Best model in the pool | 0.95 | 0.77 | 0.65 | 0.88 | 0.99 | 0.46 | 0.99 |

Table 3: On the dSprites dataset, models selected with ModelCentrality outperform those selected with UDR Lasso and UDR Spearman, for both FactorVAE and InfoGAN-CR respectively. Further, this outperforms the *hyper-parameter* tuned models supervised by the groundtruths disentangled codes reported in Table 1. Results of the model with the best FactorVAE score computed by ground truth disentangled code are also included for reference (row "best model in the pool").

| Model | FactorVAE | DCI | SAP | Explicitness | Modularity | MIG | BetaVAE |
|---|---|---|---|---|---|---|---|
| ModelCentrality | 1.00 | 0.75 | 0.77 | 0.92 | 1.00 | 0.53 | 1.00 |
| Best model in the pool | 1.00 | 0.85 | 0.89 | 0.92 | 1.00 | 0.47 | 1.00 |

Table 4: On the 3DTeapots dataset, InfoGAN-CR models selected with ModelCentrality has close performance to the model with the best groundtruth FactorVAE score and DCI score. The standard errors of ModelCentrality are less than 0.01 and we omit them in this table.

## 5. Conclusion

This work makes two contributions. First, we introduce InfoGAN-CR, a new architecture for training disentangled GANs. Next, we introduce ModelCentrality, a new framework for selecting disentangled models. Numerical results in Tables 1, 2, 3, and 4 confirm that InfoGAN-CR together with ModelCentrality achieves the best disentanglement across all metrics in the literature. This is surprising because hyper-parameter tuning in the literature is typically supervised: oracle access to the ground truth disentangled latent codes is needed. Instead, our proposed ModelCentrality is unsupervised, yet reliably selects a superior model. While ModelCentrality can be used to select both GAN and VAE based models, ModelCentrality with InfoGAN-CR improves upon ModelCentrality with other state-of-the-art methods, including VAE-based ones. Unlike other VAE-based methods, our approach seamlessly generalizes to semi-supervised settings. If we have paired examples where one latent code has been changed, e.g., a person with and without glasses, this can be readily incorporated in our architecture. Hence, one way to interpret our approach is as a self-supervised training from unsupervised data.

In addition, we experimentally find that CR substantially increases the disentanglement capabilities of InfoGAN, but does not appear to affect the state-of-the-art VAEs (Appendix F.4). Similarly, we experimentally show that the total correlation regularization, a popular technique for disentangling VAEs, do not improve disentanglement in GAN training. This suggests that disentangling VAEs and GANs require fundamentally different techniques. The proposed CR regularization could be used in any application of disentangled GANs, e.g., hierarchical image representation or reinforcement learning. Understanding this phenomenon analytically is an interesting direction for future work, and may give rise to a more general understanding of how to design regularizers for GANs as opposed to VAEs.

Another key question is to understand disentanglement in challenging datasets, compared to those studied in the literature as a benchmark. We study two such datasets. The first one studies three dimensional rotations on the 3DTeapots dataset in Appendix G.1. Existing training datasets includes only a subset of the full rotations, making disentanglement substantially easy. When training data is drawn from complete set of rotations in 3-D space, several challenges arise. The usual rotations along the three standard basis vectors do not commute, hence do not disentangle. We can find a commutative coordinate system, but it is not uniquely defined. Our preliminary experiments suggest that current state-of-the-art methods fail to learn a disentangled representation. The second one studies two dimensional polar coordinate system using a novel dataset (Circular dSprites) in Appendix H. State-of-the-art methods fail to learn the disentangled representation of the polar coordinates.

## Acknowledgements

We thank Hyunjik Kim for a productive discussion on the experimental evaluation. We thank Qian Ge for valuable discussions. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by NSF grant OCI-1053575. It used the Bridges system, which is supported by NSF award number ACI-1445606, at the Pittsburgh Supercomputing Center (PSC). This work is partially supported by funding from Siemens and AWS cloud computing credits from Amazon. Giulia Fanti acknowledges support from a Google Faculty Research Award and a JP Morgan Chase Faculty Research Award. Sewoong Oh acknowledges funding from Google Faculty Research Award, Intel Future Wireless Systems Research Award, and NSF awards CCF-1927712, CCF-1705007, IIS-1929955, and CNS-2002664.

## References

Ainsworth, S. K., Foti, N. J., and Fox, E. B. Disentangled vae representations for multi-aspect and missing data. *arXiv preprint arXiv:1806.09060*, 2018a.

Ainsworth, S. K., Foti, N. J., Lee, A. K. C., and Fox, E. B. oi-VAE: Output interpretable VAEs for nonlinear group factor analysis. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 119–128, 2018b.

Aizerman, M. A. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and remote control*, 25:821–837, 1964.

Ansari, A. F. and Soh, H. Hyperprior induced unsupervised disentanglement of latent representations. *arXiv preprint arXiv:1809.04497*, 2018.

Aumentado-Armstrong, T., Tsogkas, S., Jepson, A., and Dickinson, S. Geometric disentanglement for generative latent shape models. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 8181–8190, 2019.

Awiszus, M., Ackermann, H., and Rosenhahn, B. Learning disentangled representations via independent subspaces. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 0–0, 2019.

Bengio, Y., Courville, A., and Vincent, P. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8): 1798–1828, 2013.

Brock, A., Donahue, J., and Simonyan, K. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

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

Caselles-Dupré, H., Ortiz, M. G., and Filliat, D. Symmetry-based disentangled representation learning requires interaction with environments. In *Advances in Neural Information Processing Systems*, pp. 4608–4617, 2019.

Chen, J. and Batmanghelich, K. Weakly supervised disentanglement by pairwise similarities. *arXiv preprint arXiv:1906.01044*, 2019.

Chen, T. Q., Li, X., Grosse, R., and Duvenaud, D. Isolating sources of disentanglement in variational autoencoders. *arXiv preprint arXiv:1802.04942*, 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*, pp. 2172–2180, 2016.

Cohen, T. S. and Welling, M. Transformation properties of learned visual representations. *arXiv preprint arXiv:1412.7659*, 2014.

Creager, E., Madras, D., Jacobsen, J.-H., Weis, M. A., Swersky, K., Pitassi, T., and Zemel, R. Flexibly fair representation learning by disentanglement. *arXiv preprint arXiv:1906.02589*, 2019.

Denton, E. L. et al. Unsupervised learning of disentangled representations from video. In *Advances in neural information processing systems*, pp. 4414–4423, 2017.

Desjardins, G., Courville, A., and Bengio, Y. Disentangling factors of variation via generative entangling. *arXiv preprint arXiv:1210.5474*, 2012.

Duan, S., Watters, N., Matthey, L., Burgess, C. P., Lerchner, A., and Higgins, I. A heuristic for unsupervised model selection for variational disentangled representation learning. *arXiv preprint arXiv:1905.12614*, 2019a.

Duan, Z., Min, M. R., Li, L. E., Cai, M., Xu, Y., and Ni, B. Disentangled deep autoencoding regularization for robust image classification. *arXiv preprint arXiv:1902.11134*, 2019b.

Dupont, E. Joint-vae: Learning disentangled joint continuous and discrete representations. *arXiv preprint arXiv:1804.00104*, 2018.

Eastwood, C. and Williams, C. K. A framework for the quantitative evaluation of disentangled representations. 2018.

Esmaeili, B., Wu, H., Jain, S., Bozkurt, A., Siddharth, N., Paige, B., Brooks, D. H., Dy, J., and van de Meent, J.-W. Structured disentangled representations. *stat*, 1050:12, 2018.

Gao, S., Brekelmans, R., Ver Steeg, G., and Galstyan, A. Auto-encoding total correlation explanation. *arXiv preprint arXiv:1802.05822*, 2018.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

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 30*, pp. 6629–6640. 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. 2016.

Higgins, I., Pal, A., Rusu, A. A., Matthey, L., Burgess, C. P., Pritzel, A., Botvinick, M., Blundell, C., and Lerchner, A. Darla: Improving zero-shot transfer in reinforcement learning. *arXiv preprint arXiv:1707.08475*, 2017.

Higgins, I., Amos, D., Pfau, D., Racaniere, S., Matthey, L., Rezende, D., and Lerchner, A. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*, 2018a.

Higgins, I., Sonnerat, N., Matthey, L., Pal, A., Burgess, C. P., Bošnjak, M., Shanahan, M., Botvinick, M., Hassabis, D., and Lerchner, A. Scan: Learning hierarchical compositional visual concepts. 2018b.

Hsieh, J.-T., Liu, B., Huang, D.-A., Fei-Fei, L. F., and Niebles, J. C. Learning to decompose and disentangle representations for video prediction. In *Advances in Neural Information Processing Systems*, pp. 517–526, 2018.

Hsu, W.-N., Zhang, Y., and Glass, J. Unsupervised learning of disentangled and interpretable representations from sequential data. In *Advances in neural information processing systems*, pp. 1878–1889, 2017.

Jeon, I., Lee, W., and Kim, G. Ib-gan: Disentangled representation learning with information bottleneck gan. 2018.

Jeong, Y. and Song, H. O. Learning discrete and continuous factors of data via alternating disentanglement. *arXiv preprint arXiv:1905.09432*, 2019.

Karaletsos, T., Belongie, S., and Rätsch, G. Bayesian representation learning with oracle constraints. *arXiv preprint arXiv:1506.05011*, 2015.

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

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Kulkarni, T. D., Whitney, W. F., Kohli, P., and Tenenbaum, J. Deep convolutional inverse graphics network. In *Advances in neural information processing systems*, pp. 2539–2547, 2015.

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

Lee, H.-Y., Tseng, H.-Y., Huang, J.-B., Singh, M., and Yang, M.-H. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 35–51, 2018.

Lee, W., Kim, D., Hong, S., and Lee, H. High-fidelity synthesis with disentangled representation. *arXiv preprint arXiv:2001.04296*, 2020.

Li, Z., Tang, Y., and He, Y. Unsupervised disentangled representation learning with analogical relations. *arXiv preprint arXiv:1804.09502*, 2018.

Li, Z., Tang, Y., Li, W., and He, Y. Learning disentangled representation with pairwise independence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 4245–4252, 2019.

Lin, Z., Khetan, A., Fanti, G., and Oh, S. Pacgan: The power of two samples in generative adversarial networks. *arXiv preprint arXiv:1712.04086*, 2017.

Liu, B., Zhu, Y., Fu, Z., de Melo, G., and Elgammal, A. Oogan: Disentangling gan with one-hot sampling and orthogonal regularization. *arXiv preprint arXiv:1905.10836*, 2019.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3730–3738, 2015.

Locatello, F., Bauer, S., Lucic, M., Gelly, S., Schölkopf, B., and Bachem, O. Challenging common assumptions in the unsupervised learning of disentangled representations. *arXiv preprint arXiv:1811.12359*, 2018.

Locatello, F., Abbati, G., Rainforth, T., Bauer, S., Schölkopf, B., and Bachem, O. On the fairness of disentangled representations. In *Advances in Neural Information Processing Systems*, pp. 14584–14597, 2019a.

Locatello, F., Tschannen, M., Bauer, S., Rätsch, G., Schölkopf, B., and Bachem, O. Disentangling factors of variation using few labels. *arXiv preprint arXiv:1905.01258*, 2019b.

Lopez, R., Regier, J., Jordan, M. I., and Yosef, N. Information constraints on auto-encoding variational bayes. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31*, pp. 6117–6128. 2018.

Lorenz, D., Bereska, L., Milbich, T., and Ommer, B. Unsupervised part-based disentangling of object shape and appearance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 10955–10964, 2019.

Marx, C., Phillips, R., Friedler, S., Scheidegger, C., and Venkatasubramanian, S. Disentangling influence: Using disentangled representations to audit model predictions. In *Advances in Neural Information Processing Systems*, pp. 4498–4508, 2019.

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

Miladinović, Đ., Gondal, M. W., Schölkopf, B., Buhmann, J. M., and Bauer, S. Disentangled state space representations. *arXiv preprint arXiv:1906.03255*, 2019.

Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2018.

Moreno, P., Williams, C. K., Nash, C., and Kohli, P. Overcoming occlusion with inverse graphics. In *European Conference on Computer Vision*, pp. 170–185. Springer, 2016.

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 30*, pp. 5925–5935. 2017.

Pfister, N., Bühlmann, P., Schölkopf, B., and Peters, J. Kernel-based tests for joint independence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):5–31, 2018.

Pineau, E. and Lelarge, M. Infocatvae: Representation learning with categorical variational autoencoders. *arXiv preprint arXiv:1806.08240*, 2018.

Ridgeway, K. A survey of inductive biases for factorial representation-learning. *arXiv preprint arXiv:1612.05299*, 2016.

Ridgeway, K. and Mozer, M. C. Learning deep disentangled embeddings with the f-statistic loss. In *Advances in Neural Information Processing Systems*, pp. 185–194, 2018.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, pp. 2234–2242, 2016.

Schmidhuber, J. Learning factorial codes by predictability minimization. *Neural Computation*, 4(6):863–879, 1992.

Singh, K. K., Ojha, U., and Lee, Y. J. Finegan: Unsupervised hierarchical disentanglement for fine-grained object generation and discovery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6490–6499, 2019.

Srivastava, A., Valkov, L., Russell, C., Gutmann, M., and Sutton, C. Veegan: Reducing mode collapse in gans using implicit variational learning. *arXiv preprint arXiv:1705.07761*, 2017.

Szabó, A., Hu, Q., Portenier, T., Zwicker, M., and Favaro, P. Challenges in disentangling independent factors of variation. *arXiv preprint arXiv:1711.02245*, 2017.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Tang, Y., Salakhutdinov, R., and Hinton, G. Tensor analyzers. In *International Conference on Machine Learning*, pp. 163–171, 2013.

Tschannen, M., Bachem, O., and Lucic, M. Recent advances in autoencoder-based representation learning. *arXiv preprint arXiv:1812.05069*, 2018.

van Steenkiste, S., Locatello, F., Schmidhuber, J., and Bachem, O. Are disentangled representations helpful for abstract visual reasoning? In *Advances in Neural Information Processing Systems*, pp. 14222–14235, 2019.

Watters, N., Matthey, L., Burgess, C. P., and Lerchner, A. Spatial broadcast decoder: A simple architecture for learning disentangled representations in vaes. *arXiv preprint arXiv:1901.07017*, 2019.

Xing, X., Gao, R., Han, T., Zhu, S.-C., and Wu, Y. N. Deformable generator network: Unsupervised disentanglement of appearance and geometry. *arXiv preprint arXiv:1806.06298*, 2018.