Do Generative Models Know Disentanglement?
Contrastive Learning is All You Need

Xuanchi Ren* 1 Tao Yang* 2 Yuwang Wang 3 Wenjun Zeng 3

Abstract
Disentangled generative models are typically trained with an extra regularization term, which encourages the traversal of each latent factor to make a distinct and independent change at the cost of generation quality. When traversing the latent space of generative models trained without the disentanglement term, the generated samples show semantically meaningful change, raising the question: do generative models know disentanglement? We propose an unsupervised and model-agnostic method: Disentanglement via Contrast (DisCo) in the Variation Space. DisCo consists of: (i) a Navigator providing traversal directions in the latent space, and (ii) a Δ-Contrastor composed of two shared-weight Encoders, which encode image pairs along these directions to disentangled representations respectively, and a difference operator to map the encoded representations to the Variation Space. We propose two more key techniques for DisCo: entropy-based domination loss to make the encoded representations more disentangled and the strategy of flipping hard negatives to address directions with the same semantic meaning. By optimizing the Navigator to discover disentangled directions in the latent space and Encoders to extract disentangled representations from images with Contrastive Learning, DisCo achieves the state-of-the-art disentanglement given pre-trained non-disentangled generative models, including GAN, VAE, and Flow. Project page at https://github.com/xrenaa/DisCo.

1. Introduction
Generative models based on VAE (Kingma & Welling, 2014), GAN (Goodfellow et al., 2014) or Flow (Kingma & Dhariwal, 2018) try to model the real world by generating realistic samples from latent representations, which is a major and fundamental branch of machine learning. An important task is to enable the latent representation to disentangle the factors of variation without extra supervision and achieve controllable generation (Bengio & LeCun, 2007; Higgins et al., 2017; Eastwood & Williams, 2018). The typical approach is to introduce an extra disentanglement regularization. For example, VAE-based methods (Higgins et al., 2017; Burgess et al., 2018a; Kumar et al., 2017; Kim & Mnih, 2018a; Chen et al., 2018) use the total correlation of the latent variable distributions as the penalty, and InfoGAN-based methods (Chen et al., 2016; Lin et al., 2020) maximize the mutual information between latent factors and related observations. Usually the extra terms lead to worse generation quality for these typical disentanglement methods (Burgess et al., 2018b; Khrulkov et al., 2021).

Recently, rich attributes have been observed to emerge in the latent space of GAN purely trained for generation without any disentanglement term (Shen & Zhou, 2020; Khrulkov et al., 2021; Karras et al., 2019). These non-disentangled GAN-based methods discover semantically meaningful directions in the style space of StyleGAN (Karras et al., 2019; 2020) by analysing the distribution of the first-layer outputs (Härkönen et al., 2020) or layer weights (Shen & Zhou, 2020; Khrulkov et al., 2021). However, these GAN-based methods are limited to specific network structure and can not outperform the typical disentanglement methods.

We propose an unsupervised, model-agnostic, and interpretable disentanglement solution for the three most popular types of generative models (GAN, VAE, and Flow). Starting from the intuitive notion of disentanglement: traversal of each factor only causes a single distinct and independent variation of the generated images, given a pretrained generative model, we propose Disentanglement via Contrast (DisCo) in the Variation Space. DisCo is composed of a Navigator and a Δ-Contrastor. Specifically, as shown in Figure 1, the Navigator provides candidate traversal directions in the latent space of a generative model. Traversing along a given direction, we get an image pair and feed it to Δ-Contrastor, which consists of two shared-weight Encoders and a difference operator. The Encoders encode
the image pair to disentangled representations respectively, and the difference operator maps the disentangled representations to a sample in the Variation Space. Since the Variation Space is derived from the encoded disentangled representation explicitly, the Encoder is able to learn to extract disentangled representation without a second stage as is done in Khrulkov et al. (2021).

To optimize the Navigator to find disentangled directions in the latent space and ∆-Contrastor to extract disentangled representations for images, in the Variation Space, we pull together the samples with the same traversal directions and push away samples with different traversal directions. Furthermore, we propose two key techniques for DisCo: an entropy-based domination loss and a hard negatives flipping strategy. To encourage the encoded representations to be more disentangled, i.e., only one dimension responds when traversing along one direction in the latent space, we propose an entropy-based domination loss pushing the samples in the Variation Space towards one-hot vectors. Since the latent space of generative models is of very high dimension and may not be well-organized, there are many different directions with the same semantic meaning, resulting in hard negatives. We propose to flip these hard negatives to be positive samples for better exploring the directions.

The most similar work to ours is Voynov & Babenko (2020), which utilizes a network to reconstruct the shift (direction and scale) in the latent space from the concatenation of two generated images. Voynov & Babenko (2020) do not explicitly model the disentangled representation and rely on classification to find the disentangled directions. Contrastive Learning is better for disentanglement than classification due to: (i) the actual number of disentangled directions is usually unknown, which is similar to Contrastive Learning for retrieval (Le-Khac et al., 2020), (ii) in the latent space, there are many different directions with similar variations, and (iii) Contrastive Learning works in the representation space directly without any extra layers for classification.

We evaluate DisCo on all the three major types of generative models (GAN, VAE, and Flow) on three popular disentanglement datasets. DisCo achieves the state-of-the-art (SOTA) disentanglement performance compared to all the previous GAN-based methods and typical (InfoGAN/VAE-based) unsupervised disentanglement methods.

Our main contributions can be summarized as: (i) To our best knowledge, DisCo is the first to endow non-disentangled VAE, GAN, or Flow models with the SOTA disentanglement ability. (ii) By utilizing Contrastive Learning in the Variation Space, we can find the disentangled directions in the latent space and extract disentangled representations of images simultaneously. (iii) We propose two key techniques for DisCo: an entropy-based domination loss and a hard negatives flipping strategy.

2. Related Work

Typical unsupervised disentanglement. There have been a lot of studies on unsupervised disentangled representation learning based on VAE (Higgins et al., 2017; Burgess et al., 2018a; Kumar et al., 2017; Kim & Mnih, 2018a; Chen et al., 2018) or InfoGAN (Chen et al., 2016; Lin et al., 2020). These methods achieve disentanglement via an extra regularization, which often sacrifices the generation quality (Burgess et al., 2018b; Khrulkov et al., 2021). VAE-based methods disentangle the variations by factorizing aggregated posterior, and InfoGAN-based methods maximize the mutual information between latent factors and related observations. VAE-based methods achieve relatively good disentanglement performance, but have low-quality image generation ability. InfoGAN-based methods have a relatively high quality of generation but poor disentanglement performance. Our method supplements generative models trained without disentanglement regularization term to achieve both high-fidelity image generation and SOTA disentanglement.

Interpretable directions in the latent space. Recently, researchers have been interested in discovering the interpretable directions in the latent space of generative models without supervision, especially for GAN (Goodfellow et al., 2014; Miyato et al., 2018; Karras et al., 2020). Based on the fact that the GAN latent space often possesses semantically meaningful directions (Radford et al., 2015; Shen et al., 2020; Jahanian et al., 2020), Voynov & Babenko (2020) propose a regression-based method to explore interpretable directions in the latent space of a pretrained GAN. The following works focus on extracting the directions from a specific layer of GANs. Härkönen et al. (2020) searches for important and meaningful directions by performing PCA in the style space of StyleGAN (Karras et al., 2019; 2020). Shen & Zhou (2020) propose to use the singular vectors of the first layer of a generator as the interpretable directions, and Khrulkov et al. (2021) extend this method to the intermediate layers by Jacobian matrix. Khrulkov et al. (2021) also bridge unsupervised disentanglement and this line of works by training an extra encoder. However, all these methods can not outperform the typical disentanglement methods.

Contrastive learning. Contrastive Learning gains popularity due to its effectiveness in representation learning (He et al., 2020; Grill et al., 2020; van den Oord et al., 2018; Henaff, 2020; Li et al., 2020; Chen et al., 2020). Typically, contrastive approaches bring representation of different views of the same image (positive pairs) closer, and push representations of views from different images (negative pairs) apart using instance-level classification with contrastive loss. In this work, we focus on the variations of representations and achieve SOTA disentanglement with Contrastive Learning.
Figure 1. Overview of DisCo. DisCo consists of: (i) a Navigator exploring traversal directions in the latent space of a given pretrained generative model, (ii) a Δ-Contrastor encoding traversed images into the Variation Space, where we utilize Contrastive Learning. Samples in the Variation Space correspond to the image variations along the directions provided by the Navigator labeled with different colors, respectively. Δ-Contrastor includes two shared-weight encoders encoding traversed image pair to disentangle representations respectively, and a difference operator getting the variation between the encoded representations. The Generative Model is fixed, and the Navigator and Encoders are learnable marked with grey color.

3. Disentanglement via Contrast

3.1. Intuition on DisCo

The disentangled representation is proposed by Bengio et al. (2013); Higgins et al. (2017); Eastwood & Williams (2018) to disentangle the variation into independent factors. For a disentangled generative model, each factor controls a distinct and interpretable variation of the generated images. Those disentangled generative models can be achieved by training with an extra disentanglement term and a hyper-parameter to balance disentanglement and image generation. For generative models trained without a disentanglement term, traversal along different directions in the latent space results in different kinds of image variations. The traversal directions in the latent spaces are disentangled only if the corresponding image variations are independent and distinct from each other. Inspired by this intuitive notion of disentanglement, we propose Disentanglement via Contrast (DisCo), which is an unsupervised and model-agnostic method to find the disentangled directions in the latent space and extract disentangled representations from pretrained generative models simultaneously.

DisCo makes the general pretrained generative models (VAE, GAN, and FFlow) disentangled by discovering independent and distinct directions in the latent space, and the overview of our method is shown in Figure 1. Given a pretrained generative model $G: \mathcal{Z} \rightarrow \mathcal{I}$, where $\mathcal{Z}$ denotes the latent space and $\mathcal{I}$ denotes the image space, the working flow of DisCo is as follows: 1) A Navigator $A$ provides candidate traversal directions in the latent space $\mathcal{Z}$. 2) Image pairs $G(z)$, $G(z')$ are generated from $z, z' \in \mathcal{Z}$, and $z' = z + A(d, \varepsilon)$, where $d$ denotes the index of the direction, and $\varepsilon$ denotes the shift scalar. 3) The Δ-Contrastor is composed of two shared-weight Encoders $E$ to encode the image pair to a sample $v \in \mathcal{V}$ as $v(z, d, \varepsilon) = |E(G(z + A(d, \varepsilon))) - E(G(z))|$, where $\mathcal{V}$ denotes the Variation Space.

Then we apply Contrastive Learning in $\mathcal{V}$ to optimize the Encoder $A$ to find the disentangled directions in latent space $\mathcal{Z}$, and the Encoder $E$ to extract disentangled representations simultaneously.

3.2. Design of DisCo

We present the design details of DisCo including: (i) the collection of query set $Q$, positive key set $K^+$ and negative key set $K^-$ in Variation Space, (ii) the formulation of the Constrastive Loss.

To generate $Q = \{q_i\}^B_{i=1}$, $K^+ = \{k^+_i\}^N_{i=1}$ and $K^- = \{k^-_i\}^M_{i=1}$, where $B$ is the number of query samples, $N$ is the number of positive samples, and $M$ is the number of negative samples, we need to design the sampling strategy of $z, d, \varepsilon$ for each sample to ensure that the corresponding samples from $Q$ and $K^+$ have the same direction and the generated samples can cover a variation range as large as possible. We first sample the direction index $d$ for $Q$ and $K^+$ from a discrete uniform distribution $\mathcal{U}\{1, D\}$, where $D$ is the total number of directions. For $K^-$, we sample the directions $\{d^-_i\}^M_{i=1}$ from $\mathcal{U}\{1, \mathcal{D}\} \setminus \{d\}$ for $\{k^-_i\}^M_{i=1}$, respectively and independently. For $Q$, $K^+$ and $K^-$, we sample $\{z_i\}^B_{i=1}$, $\{z^+_i\}^N_{i=1}$; $\{z^-_i\}^M_{i=1}$ from latent space $\mathcal{Z}$, and sample $\{\varepsilon_i\}^B_{i=1}$, $\{\varepsilon^+_i\}^N_{i=1}$; $\{\varepsilon^-_i\}^M_{i=1}$ from a continuous uniform distribution $\mathcal{U}[-\varepsilon, \varepsilon]$ respectively and independently. For $A$, we adopt the same form as Voinov & Babenko (2020), $A(d, \varepsilon) = A(\varepsilon e_d)$, where $e_d$ denotes an axis-aligned unit vector $(0, ..., 1_d, ..., 0)$. $A$ can be a linear operator or a non-linear multi-layer perceptron (MLP).
Each sample $q_i \in Q$ is generated as:

$$q_i = |E(G(zi) + A(\varepsilon_i e_d))) - E(G(zi))|. (1)$$

Each sample $k_i^+ \in \mathcal{K}^+$ is generated as:

$$k_i^+ = |E(G(z_i^+ + A(\varepsilon_i^+ e_d))) - E(G(z_i^+))|. (2)$$

Each sample $k_i^- \in \mathcal{K}^-$ is generated as:

$$k_i^- = |E(G(z_i^- + A(\varepsilon_i^- e_d^-))) - E(G(z_i^-))|. (3)$$

For $Q$, $\mathcal{K}^+$ and $\mathcal{K}^-$, we normalize each sample to a unit vector to eliminate the impact caused by different shift scalars.

We follow InfoNCE (van den Oord et al., 2018) to use the dot product to measure the similarity, and the contrastive loss function is:

$$L_{NCE} = -\frac{1}{|B|} \sum_{i=1}^{B} \log \sum_{j=1}^{N} \exp(q_i \cdot k_j^+ / \tau) \sum_{m=1}^{M} \exp(q_i \cdot k_m^- / \tau), (4)$$

where $\tau$ is a temperature hyper-parameter. As mentioned in InfoNCE (van den Oord et al., 2018), BCELoss $L_{logits}$ is a lower bound of NCELoss $L_{NCE}$. We follow them to use BCELoss $L_{logits}$ to substitute the NCELoss for better optimization:

$$L_{logits} = \frac{1}{|B|} \sum_{i=1}^{B} (l_i^- + l_i^+),$$

$$l_i^+ = \sum_{j=1}^{N} \log \sigma(q_i \cdot k_j^+ / \tau),$$

$$l_i^- = \sum_{m=1}^{M} \log(1 - \sigma(q_i \cdot k_m^- / \tau)),$$ (5)

where $\sigma$ denotes the sigmoid function, $l_i^+$ denotes the part for positive samples, and $l_i^-$ denotes the part for the negative ones.

### 3.3. Key Techniques for DisCo

Targeting for the disentanglement task, we propose two more key techniques for DisCo to achieve better disentanglement.

**Entropy-based domination loss.** By optimizing contrastive loss, the Navigator $A$ is optimized to find the disentangled directions in latent space, and Encoder $E$ is optimized to extract disentangled representations of images. To further make the encoded representations more disentangled, i.e., when traversing along one disentangled direction, only one dimension of the encoded representation should respond, and the corresponding sample in the Variation Space should be one-hot. We thus propose an entropy-based domination loss to encourage the samples to be one-hot.

To implement the entropy-based domination loss, we first get the mean $c$ of the query set $Q$ and the positive key set $\mathcal{K}^+$ as

$$c = -\frac{1}{|B + N|} \left( \sum_{i=1}^{B} q_i + \sum_{i=1}^{N} k_i^+ \right). (6)$$

We then compute the probability $p_i$ as

$$p_i = \frac{\exp(c(i))}{\sum_{j=1}^{n} \exp(c(j))}, (7)$$

where $c(i)$ is the $i$-th element of $c$ and $n$ is the total number of dimensions of $c$. The entropy-based domination loss $L_{ed}$ is calculated as

$$L_{ed} = \frac{1}{|n|} \sum_{i=1}^{n} p_i \log(p_i), (8)$$

**Hard negatives flipping.** Since the latent space of the generative models is a high-dimension complex manifold, there are many different directions carrying the same semantic meaning as demonstrated in Section 4.4. These directions with the same semantic meaning result in hard negatives during the optimization of Contrastive Loss. The hard negatives here are different from the hard negatives in the works of self-supervised representation learning (He et al., 2020; Coskun et al., 2018), where they have reliable annotations of the samples. Here, our hard negatives are more likely to be “false” negatives, and we choose to flip these hard negatives into positives. Specifically, we use a threshold $T$ to identify the hard negative samples, and use their similarity to the queries as the pseudo-labels for them:

$$\hat{l}_i = \sum_{\alpha_{ij} < T} \log(1 - \sigma(\alpha_{ij})) + \sum_{\alpha_{ij} \geq T} \alpha_{ij} \log(\sigma(\alpha_{ij})), (9)$$

where $\hat{l}_i$ denotes the modified $l_i^-$, and $\alpha_{ij} = q_i \cdot k_j^- / \tau$. Therefore, the modified final BCELoss is:

$$L_{logits-f} = \frac{1}{|B|} \sum_{i=1}^{B} (l_i^+ + \hat{l}_i^-). (10)$$

**Full objective.** With the above two techniques, the full objective is:

$$L = L_{logits-f} + \lambda L_{ed}, (11)$$

where $\lambda$ is the weighting hyper-parameter for entropy-based domination loss $L_{ed}$. 
4. Experiment

4.1. Experimental Setup

Datasets. We consider the following popular datasets in the disentanglement areas: Shapes3D (Kim & Mnih, 2018b) containing 480,000 images with 6 ground truth factors, MPI3D (Gondal et al., 2019) containing 1,036,800 robotic arm images with 7 ground truth factors, and Cars3D (Reed et al., 2015) consisting of 17,568 images with 3 ground truth factors. In all the experiments, images are resized to the 64x64 resolution.

Pretrained generative models. For GAN, we use the StyleGAN2 model (Karras et al., 2020). For VAE, we use a simple structure with convolutions. For Flow, we use Glow (Kingma & Dhariwal, 2018).

Baseline. For the typical disentanglement baselines, we choose FactorVAE (Kim & Mnih, 2018a), β-TCVAE (Chen et al., 2018) and InfoGAN-CR (Lin et al., 2020). For GAN-based methods that extract disentangled representations from pretrained GANs, we consider several recent methods: GANspace (GS) (Härkönen et al., 2020), LatentDiscovery (LD) (Voynov & Babenko, 2020), ClosedForm (CF) (Shen & Zhou, 2020) and DeepSpectral (DS) (Khrulkov et al., 2021). For these methods, we follow Khrulkov et al. (2021) to train an additional encoder for evaluation. We are the first to extract disentangled representations from pretrained VAE and Flow, so we extend LD to VAE and Flow as a baseline.

Disentanglement metrics. For the disentanglement metrics, we mainly consider two representative ones: the Mutual Information Gap (MIG) (Chen et al., 2018) and the Disentanglement metric (DCI) (Eastwood & Williams, 2018). MIG measures, for each factor of variation, the normalized gap between the top two entries of the pairwise mutual information matrix, which requires each factor to be only perturbed by changes of a single dimension of representation. DCI measures, for each dimension of representation, the importance of predicting a single dominant factor, which requires each dimension only to encode the information of a single dominant factor. With these two metrics, we can evaluate the entanglement in terms of both representation and factors. We also provide results for β-VAE score (Higgins et al., 2017) and FactorVAE score (Kim & Mnih, 2018a) in Appendix.

Random seeds. For StyleGAN2 and VAE, we train five models of different random seeds on all three datasets. For each model of StyleGAN2 and VAE, we also use five random seeds to train our method. For Glow, we only train a single model and use one random seed to train our method, limited by GPU resource. The random seed setting for the baselines is presented in Appendix.

Figure 2. Violin plots (Hintze & Nelson, 1998) of MIG and DCI scores (higher is better) for various methods on the Shapes3D dataset (white dots indicate means). Ours is DisCo on pretrained GAN and achieves the best performance. Each method has 25 runs, and the variance is due to random seeds.

Figure 3. Examples of disentangled directions for StyleGAN2 on FFHQ.
Do Generative Models Know Disentanglement? Contrastive Learning is All You Need

![Figure 4. Visualization of the learned disentangled space. We feed the images traversing the ground truth factor space (the three most significant factors are considered) to the trained encoders and plot the derived representations of the corresponding dimensions in 3D space. The localization of each point is the disentangled representation of the corresponding image. And an ideal result is that all the points form a cube and color variation is continuous. Our latent space is less distorted and aligns the axes better than the baseline methods, which implies better disentanglement.](image)

![Figure 5. Examples of disentangled directions for Spectral Norm GAN on Anime Faces and MNIST.](image)

**Figure 4.** Visualization of the learned disentangled space. We feed the images traversing the ground truth factor space (the three most significant factors are considered) to the trained encoders and plot the derived representations of the corresponding dimensions in 3D space. The localization of each point is the disentangled representation of the corresponding image. And an ideal result is that all the points form a cube and color variation is continuous. Our latent space is less distorted and aligns the axes better than the baseline methods, which implies better disentanglement.

**Analysis of the learned disentangled representations.** We feed the images traversing the three most significant factors (wall color, floor color, and object color) of Shapes3d into the encoders inside $\Delta$-Contrastor and plot the corresponding dimensions of the encoded representations to visualize the learned disentangled space. We only consider three baselines that have relatively higher MIG and DCI: CF, DS, LD. As Figure 4 shows, the points in the latent space of CF and DS are not well organized, and the latent space of all the three baselines are not well aligned with the axes, especially for LD. DisCo learns a well-aligned and well-organized latent space, which signifies a better disentanglement.

**4.3. Qualitative Evaluations**

Besides disentanglement datasets, DisCo can also discover disentangled and meaningful directions for real-world datasets, e.g. FFHQ (Karras et al., 2019), as shown in Figure 3. These directions are useful for controllable image generation. Besides StyleGAN2, we also provide results of Spectral Norm GAN (Miyato et al., 2018) on MNIST (Le-

Cun et al., 2010) and Anime Face (Jin et al., 2017). As shown in Figure 5, DisCo can be well generalized to other types of GAN. For qualitative results on more datasets and more types of generative models, please refer to Appendix.

**4.4. Ablation Study**

In this section, we perform ablation study of DisCo only on GAN, limited by the space. For the experiments, we use the Shapes3D dataset, and the random seed is fixed.

**Choice of latent space.** For style–based GANs (Karras et al., 2019; 2020), there is a style space $W$, which is the output of style network (MLP) whose input is a random latent space $Z$. As demonstrated in Karras et al. (2019), $W$ is more interpretable than $Z$. We conduct experiments on $W$ and $Z$ respectively to see how the degree of disentanglement of the latent space influences the performance. As shown in Table 2, DisCo on $W$ is better, indicating that the better the latent space is organized, the better disentanglement DisCo can achieve.

**Choices of $A$.** Following the setting of Voynov & Babenko (2020), we mainly consider three options of $A$:

- $A$ is a linear operator with all matrix columns having a unit length;
- $A$ is a linear operator with orthonormal matrix columns;
- $A$ is a non-linear operator of 3 fully-connected layers.

The results are shown in Table 2. For latent spaces $W$ and $Z$, $A$ with unit-norm columns achieves nearly the best performance in terms of MIG and DCI scores. Compared to $A$ with orthonormal matrix columns, using $A$ with unit-norm columns is more expressive with less constraints. $A$ composed of 3 fully-connected layers performs poorly, indicating the disentangled directions of the latent space $W$ of StyleGAN is nearly linear. Another possible reason is that $A$ is global without considering the latent code $z$. A non-linear operator is more suitable for a local navigator.
Do Generative Models Know Disentanglement? Contrastive Learning is All You Need

Table 1. Comparisons of the MIG and DCI disentanglement metrics (mean ± variance). A higher mean indicates a better performance. DisCo can extract disentangled representations from all three generative models, and DisCo on GAN achieves the highest score in almost all the cases, compared to all the baselines. All the cells except for Flow are averaged results over 25 runs. More results and details about the settings are presented in Appendix.

Table 2. Ablation study of DisCo on the choice of the latent spaces, the choice of the types of $A$, and our proposed techniques.

Table 2. Ablation study of DisCo on the choice of the latent spaces, the choice of the types of $A$, and our proposed techniques.

**Table 2.** Comparisons of the MIG and DCI disentanglement metrics (mean ± variance). A higher mean indicates a better performance. DisCo can extract disentangled representations from all three generative models, and DisCo on GAN achieves the highest score in almost all the cases, compared to all the baselines. All the cells except for Flow are averaged results over 25 runs. More results and details about the settings are presented in Appendix.

For such a much more complex local and non-linear setting, more inductive bias or supervision should be introduced.

**Entropy-based domination loss.** Here, we verify the effectiveness of entropy-based domination loss $\mathcal{L}_{cd}$ for disentanglement. For a desirable disentangled representation, one semantic meaning corresponds to one dimension. As shown in Table 2, $\mathcal{L}_{cd}$ can improve the performance by a large margin. We also visualize the Variation Space to further demonstrate the effectiveness of our proposed loss in Figure 7. Adding the domination loss makes the samples in the Variation Space to be one-hot, which is desirable for disentanglement.

**Hard negatives flipping.** We run our DisCo with or without the hard negatives flipping strategy to study its influence. As shown in Table 2, flipping hard negatives can improve the disentanglement ability of DisCo. The reason is that the hard negatives have the same semantics as the positive samples, as shown in Figure 8. In this case, treating them as the hard negatives does not make sense. Flipping them with pseudo-labels can make the optimization of contrastive learning easier.

**Hyperparameter N & M.** We run DisCo with different ratios of $N : M$ with a fixed total amount 96 and different total amount $N + M$ with a fixed ratio $1 : 2$ to study their impacts. As shown in Figure 6 (a), the best ratio is $N : M = 32 : 64 = 1 : 2$, as the red line (MIG) in the figure shows that larger or smaller ratios will hurt DisCo, which indicates DisCo requires a balance between $N$ and $M$. As the blue line (DCI) shows, DCI drops significantly when the ratio increases, suggesting that the percentage of negatives are important for DisCo. DisCo can pull together positives better with the higher ratio of negatives and the variations can be well learned with enough positives. As shown in Figure 6 (b), the total amount of samples has slight impact on DisCo.
Do Generative Models Know Disentanglement? Contrastive Learning is All You Need

Figure 6. Study on numbers of positive and negative samples. The balance between positive and negative samples is crucial for DisCo.

Figure 7. Visualization of the variation of the encoded disentangled representations caused by the change of a single ground truth factor.

Table 3. Ablation on Contrast v.s. Classification and Concatenation (Concat) v.s. Variation.

| Concat | Variation | Contrast | Classification | MIG  | DCI  |
|--------|-----------|----------|-----------------|------|------|
| ✓      | ✓         | ✓        |                 | 0.023| 0.225|
| ✓      | ✓         | ✓        |                 | 0.562| 0.736|
| ✓      | ✓         | ✓        |                 | 0.012| 0.138|
| ✓      | ✓         | ✓        |                 | 0.002| 0.452|

Contrast v.s. Classification. To verify the effectiveness of Contrastive Learning, we substitute it with classification by adopting an additional linear layer to predict the index of directions. As Table 3 shows, Contrastive Learning outperforms Classification.

Concatenation v.s. Variation. We further demonstrate that the Variation Space is crucial for DisCo. By replacing the difference operator with concatenation, the performance drops significantly (Table 3), indicating that the encoded representation is not well disentangled. On the other hand, the Contrastive Learning in the Variation Space optimizes the Encoder to extract disentangled representations of images.

4.5. Analysis of Different Generative Models

As shown in Table 1, DisCo can be well generalized to different generative models (GAN, VAE, and Flow). DisCo on GAN and VAE can achieve relative good performance, while DisCo on Flow is not as good. The possible reason is similar to the choice of latent space of GAN. We assume the disentangled directions are global linear and thus use a linear navigator. In contrast to GAN and VAE, we suppose that Flow may not conform to this assumption well. Furthermore, Flow has the problems of high GPU cost and unstable training, which limit us to do further exploration.

5. Conclusion

In this paper, we present an unsupervised and model-agnostic method DisCo, which can discover the disentangled directions in the latent space of pretrained generative models and extract disentangled representations of images. We propose an entropy-based domination loss and a hard negatives flipping strategy to achieve better disentanglement. DisCo beats typical unsupervised disentanglement methods and maintains high image quality as well. We pinpoint a new direction that Contrastive Learning can be well applied to extract disentangled representation from pretrained generative models. For future work, extending DisCo to the existing VAE-based disentanglement framework is an exciting direction.
References

Bengio, Y. and LeCun, Y. Scaling learning algorithms towards AI. In Large Scale Kernel Machines. MIT Press, 2007.

Bengio, Y., Courville, A. C., and Vincent, P. Representation learning: A review and new perspectives. PAMI, 2013.

Burgess, C. P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., and Lerchner, A. Understanding disentangling in $\beta$-vae. CoRR, abs/1804.03599, 2018a.

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, 2018b.

Chen, R. T., Li, X., Grosse, R. B., and Duvenaud, D. K. Isolating sources of disentanglement in variational autoencoders. In NeurIPS, 2018.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. E. A simple framework for contrastive learning of visual representations. In ICML, 2020.

Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., and Abbeel, P. Infogan: interpretable representation learning by information maximizing generative adversarial nets. In NeurIPS, 2016.

Coskun, H., Tan, D. J., Conjeti, S., Navab, N., and Tombari, F. Human motion analysis with deep metric learning. In ECCV, 2018.

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

Gondal, M. W., Wuthrich, M., Miladinovic, 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. In NeurIPS, 2019.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. C., and Bengio, Y. Generative adversarial networks. CoRR, abs/1406.2661, 2014. URL http://arxiv.org/abs/1406.2661.

Grill, J., Strub, F., Altché, F., Tallec, C., Richemond, P. H., Buchatskaya, E., Doersch, C., Pires, B. A., Guo, Z., Azar, M. G., Piot, B., Kavukcuoglu, K., Munos, R., and Valko, M. Bootstrap your own latent - A new approach to self-supervised learning. In NeurIPS, 2020.

Härkönen, E., Hertzmann, A., Lehtinen, J., and Paris, S. Ganspace: Discovering interpretable GAN controls. In NeurIPS, 2020.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In CVPR, 2020.

Hénaff, O. J. Data-efficient image recognition with contrastive predictive coding. In ICML, 2020.

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.

Hintze, J. L. and Nelson, R. D. Violin plots: a box plot-density trace synergism. The American Statistician, 52 (2):181–184, 1998.

Jahanian, A., Chai, L., and Isola, P. On the “steerability” of generative adversarial networks. In ICLR, 2020.

Jin, Y., Zhang, J., Li, M., Tian, Y., and Zhu, H. Towards the high-quality anime characters generation with generative adversarial networks. In Proceedings of the Machine Learning for Creativity and Design Workshop at NeurIPS, 2017.

Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. In CVPR, 2019.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. Analyzing and improving the image quality of stylegan, 2021. URL https://openreview.net/forum?id=VCAXR34cp59.

Kim, H. and Mnih, A. Disentangling by factorising. In ICML, 2018a.

Kim, H. and Mnih, A. Disentangling by factorising. In ICML, 2018b.

Kingma, D. P. and Dhariwal, P. Glow: Generative flow with invertible 1x1 convolutions. In NeurIPS, 2018.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. In ICLR, 2014.

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

Le-Khac, P. H., Healy, G., and Smeaton, A. F. Contrastive representation learning: A framework and review. IEEE Access, 2020.
LeCun, Y., Cortes, C., and Burges, C. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

Li, J., Zhou, P., Xiong, C., Socher, R., and Hoi, S. C. H. Prototypical contrastive learning of unsupervised representations. *CoRR*, abs/2005.04966, 2020. URL https://arxiv.org/abs/2005.04966.

Lin, Z., Thekumparampil, K., Fanti, G., and Oh, S. Infogancr and modelcentrality: Self-supervised model training and selection for disentangling gans. In *ICML*, 2020.

Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. Spectral normalization for generative adversarial networks. In *ICLR*, 2018.

Radford, A., Metz, L., and Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. In *ICLR*, 2015.

Reed, S. E., Zhang, Y., Zhang, Y., and Lee, H. Deep visual analogy-making. In *NeurIPS*, 2015.

Shen, Y. and Zhou, B. Closed-form factorization of latent semantics in gans. 2020. URL https://arxiv.org/abs/2007.06600.

Shen, Y., Gu, J., Tang, X., and Zhou, B. Interpreting the latent space of gans for semantic face editing. In *CVPR*, 2020.

van den Oord, A., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018. URL http://arxiv.org/abs/1807.03748.

Voynov, A. and Babenko, A. Unsupervised discovery of interpretable directions in the GAN latent space. In *ICML*, 2020.