An Empirical Study on Disentanglement of Negative-free Contrastive Learning

Anonymous Author(s)
Affiliation
Address
email

Abstract

Negative-free contrastive learning has attracted a lot of attention with simplicity and impressive performances for large-scale pretraining. But its disentanglement property remains unexplored. In this paper, we take different negative-free contrastive learning methods to study the disentanglement property of this genre of self-supervised methods empirically. We find the existing disentanglement metrics fail to make meaningful measurements for the high-dimensional representation model so we propose a new disentanglement metric based on Mutual Information between representation and data factors. With the proposed metric, we benchmark the disentanglement property of negative-free contrastive learning for the first time, on both popular synthetic datasets and a real-world dataset CelebA. Our study shows that the investigated methods can learn a well-disentangled subset of representation. We extend the study of the disentangled representation learning to high-dimensional representation space and negative-free contrastive learning for the first time.

1 Introduction

Learning a disentangled representation is a long-desired goal in the deep learning community (5; 37; 13; 4; 42; 28; 46). A disentangled representation matches how humans understand the world and allows us to use much fewer labels to learn challenging downstream tasks (48). Previous disentangled learning research is usually limited to generative models(24; 16; 23; 12; 9; 31), simple synthetic datasets (36; 6) and low representation dimensions, e.g., no more than 20-d.

In contrast to generative methods, contrastive learning is a class of discriminative methods. It is trained in a self-supervised manner by pulling the representation of two augmentations of the same image close. Recent contrastive methods have achieved advanced performance on image pretraining tasks and require high representation dimensions, e.g., usually at least 1000-d.

Given the success of contrastive learning, it remains unknown whether it can lead to disentangled representations. Recently, some works (49; 53) reveal that contrastive learning can approximately invert the data generation process and this makes the learned representation have identifiability, which is related to the disentanglement property. However, all these advances necessarily rely on the contrast provided by negative samples. The disentanglement property of contrastive learning without negatives, or "non-contrastive self-supervised learning", remains unexplored.

Therefore, in this work, we study the disentanglement property of contrastive learning without using negatives. We study this empirically by introducing it into disentangled representation learning benchmarks. Given the fact that self-supervised contrastive learning methods require a high dimension of latent representation, This continues to bring a derivative study of quantitative disentanglement measurement in a high-dimensional representation space, which remains blank as well. We find though multiple disentanglement metrics (16; 23; 7; 11; 27) are demonstrated well aligned in low-dimensional space (34), they do not agree anymore in high-dimensional space. Moreover, the design limitation of them makes the measurement even not meaningful anymore. We thus propose

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

a new disentanglement evaluation metric that is robust to representation dimension change and thus adaptable to a high-dimensional representation model. The new metric is named as "**M**utual information based **E**ntroy **D**isentanglement" and MED for short.

With the proposed metric, we find that negative-free contrastive learning methods can achieve good disentanglement in a subset of latent dimensions. Given that the recent disentanglement study is limited to simple synthetic datasets but contrastive learning is especially powerful for complicated and large-scale datasets, we also extend the quantitative benchmarking to a real-world dataset CelebA (33). On CelebA, existing low-dimensional generative disentangled learning methods can not learn good representation anymore, suggesting the gap between current disentangled learning research and the real-world data complexity.

To summarize, our contributions in the work are three-folded:

1. We find that existing disentanglement metrics fail to extend to high-dimensional representation space and propose a new metric MED to extend this area to high-dimensional space.

2. We extend the study of disentangled representation learning to real-world complicated datasets and high-dimensional representation space, revealing the gap between current disentangled representation learning and real-world data complexity.

3. We empirically study the disentanglement property of contrastive learning without negatives. We find it can learn a well-disentangled subspace of latent representation.

2 Related Works

Disentangled Representation Learning. Disentangled representation aims to represent a human interpretable pattern (5; 9; 26), to make the downstream tasks learned more easily (48) and generalize better (1). There are two lines of work related to this problem, i.e., the Independent Component Analysis (ICA) and the Disentangled Deep Representation Learning. ICA (18) usually assumes that the pattern of noise (17; 21) or some additional auxiliary variables (19; 22) can be observed. On the other hand, deep representation learning makes no explicit assumption on the noise distribution or representation prior and usually emphasizes unsupervised learning. Under this setting, deep representation learning is usually based on deep generative models such as VAE-based methods (16; 23; 7; 27) and Generative Adversarial Networks (GAN) (12; 9). These two lines of study focus on different but relevant (25) aspects of a learned representation. In this paper, we focus on empirically studying the disentanglement property of negative-free contrastive learning methods to introduce them into the scope of disentangled representation learning.

Disentanglement Metrics. The measurement also shows the difference between ICA and Disentangled Deep Representation Learning. ICA aims to achieve good identifiability and uses Mean Correlation Coefficient (MCC) as the common metric. On the other hand, the metrics used in the deep disentangled representation learning community are very diverse. DisLib (34) (ICML 2019 best paper) makes an intensive empirical study of a stack of deep disentangled representation learning methods and summarizes six popular disentanglement metrics, i.e., DCI (11), SAP (27), MIG (7), BetaVAE score (16), FactorVAE score (23) and Modularity (41). DisLib finds that the five metrics except for Modularity have good agreement in evaluating disentanglement quality. However, all of the metrics are evaluated on low-dimensional latent space, which is around 10 dimensions. We find severe problems when applying those metrics to high-dimensional latent spaces where they show significant disagreement. Our proposed metric, i.e., Mutual Information based Entropy Disentanglement, is designed to be more robust to high dimensional latent spaces and is a proper tool to evaluate the disentanglement of contrastive learning methods.

Contrastive Learning. Contrastive learning creates "views" by augmentations over images. Views of the same image serve as positives, and views of other images as negatives. This setup is also known as examplar classification (10), or instance discrimination (50). Recently, some works try to understand contrastive learning either theoretically (49; 2; 45; 44; 30) or empirically (43; 52; 38). Zimmermann et al. suggests that the contrastive method inverts a data generation process, which is related to the study of disentanglement property. This conclusion is based on the analysis of Wang and Isola where infinite negative samples are required. Moreover, empirical evidence supporting these theorems is inadequate and fails in real-world practice where some assumptions are violated (e.g., augmented data distribution). On the other hand, the disentanglement property of negative-free contrastive learning (40; 8; 51) remains unexplored, either empirically or theoretically. Thus, we

aim to empirically study whether negative-free self-supervised learning can lead to disentangled representations in this paper.

3 Method and Metric

In Section 3.1, we review the methods of negative-free contrastive learning. Then we introduce our proposed MI-based Entropy Disentanglement score (MED) for evaluating disentanglement in high-dimensional representation space in Section 3.2 and its variant to evaluate disentanglement of a subspace of representation in Section 3.3.

3.1 Negative-free Contrastive Learning Methods

There are multiple negative-free contrastive learning methods proposed. They share critical characteristics that they do not use the contrast between positive samples and negative samples to encourage discriminative representations. They all generate positive "views" of data by applying different augmentations to the same input data and a pair of views is respectively forward into two steams of networks. These two networks have a latent encoder, namely the representation network, of the same architecture. However, recent works on this line, such as BYOL (40), SimSiam (8) and Barlow Twins (51), have their own unique designs.

As the first and the most classic method in this line, BYOL (40) discovers that self-supervised learning can avoid trivial solutions, i.e. "model collapse", even without using negative samples to provide contrast. The key of BYOL is to add a predictor layer following the commonly adopted "encoder-projector" network of contrastive learning methods (15). This provides additional asymmetry. Besides this, BYOL also proposes the stop-gradient strategy in the stream of the target network but its weights are still updated by momentum from the online network. As a follow-up, SimSiam (8) continues to remove the momentum update from BYOL and proves that "predictor+stop-gradient" is enough for self-supervised learning to learn non-trivial representation. More recently, Barlow Twins (51) provides a new perspective where even the predictor or the trick of stop-gradient is not necessary anymore. It moves the target of learning from samples to the embedding space. The key idea is to make the empirical cross-correlation matrix among different views close to the identity. This encourages the embedding vectors of views, which are distorted versions of the same sample, to be similar while minimizing the redundancy between these vectors.

Positive-negative contrast in self-supervised learning is considered useful to encourage learned representation to have uniformity on a hyper-sphere (49). When negative views are not available anymore, the property of the learned representation remains a mystery, especially from the perspective of disentanglement we focus on in this work. Therefore, we aim to take the mentioned negative-free contrastive learning methods into an empirical study to reveal this property of interest.

3.2 Mutual Information based Entropy Disentanglement

Typical contrastive learning methods need a high dimensional representation space, or "latent space", to train well. However, the previous study of disentangled representation learning only deals with low-dimensional representation space. For example, in Locatello et al., the latent space dimension is set to 10. So existing disentanglement metrics (see Appendix C in the supplementary materials for details) are designed for low-dimensional representation model and have intrinsic flaws in extending disentanglement evaluation to high-dimensional space. To be precise, we have observations as below:

- Metrics based on learnable classifiers, such as **BetaVAE score** and **FactorVAE score**, allow unfair advantages to high-dimensional model whose redundant parameters can trick the classifier more easily. For example, a randomly initialized 1000-d model could reach a FactorVAE score of 61.4 on dSprites, close to many well-trained 10-d VAE-based models' scores (see Table 4 in Appendix).
- Metrics taking only one or two dimensions into score calculation, such as **SAP** and **MIG**, are biased to representations of different dimensions. Because a higher dimension makes it harder for an informative dimension to stand out and enjoy a large informativeness gap over other dimensions.
- **DCI Disentanglement score** uses a learnable regressor to score the importance of each latent dimension to each data factor. The learnable regressor, such as Gradient Boosting Tree (GBT), encourages sparsity in the output importance matrix (see Figure 6 in Appendix). So it also gives an advantage to high-dimensional models, making it unfair to compare models of different latent dimensions. Moreover, the construction of regressors is time-intensive in high-dimension space. For example, it usually takes hours to evaluate a 1000-d representation model by DCI using GBT.

These flaws will be demonstrated by our experiments. More importantly, when the dimension of the representation model is high, these existing metrics do not agree with each other anymore. This proves that they are no longer eligible to make meaningful disentanglement measurements in the high-dimensional representation space.

Given (1) the bias and limitations of existing disentanglement metrics, (2) the nature of contrastive learning to require high-dimensional representation, and (3) our objective to evaluate its disentanglement, we propose a new disentanglement metric for high dimensional latent spaces, which we name as "**M**utual Information based **E**ntropy **D**isentanglement", or MED in short. The calculation of MED is based on mutual information between latent dimensions and dataset ground truth factors. Given a dataset generated by K ground truth factors $\boldsymbol{v} \in \mathbb{R}^K$ and a representation vector $\boldsymbol{c} \in \mathbb{R}^D$, we construct an importance matrix $R \in \mathbb{R}^{D \times K}$ defined by

$$R_{ij} = I(\boldsymbol{c}_i, \boldsymbol{v}_j) / \sum_{d=0}^{D-1} I(\boldsymbol{c}_d, \boldsymbol{v}_j), \tag{1}$$

where $I(\boldsymbol{c}_i, \boldsymbol{v}_j)$ denotes the mutual information between the ith latent dimension $\boldsymbol{c}_i$ and the jth ground truth factor $\boldsymbol{v}_j$. Here, each row denotes a representation dimension and each column represents a ground truth factor. We normalize the mutual information by columns, such that an entry in the matrix indicates the relative importance of one dimension over all dimensions. This normalization is necessary before computing the disentanglement level among factors, since different dimensions may have different entropy.

After normalizing over the columns, we evaluate the contribution of a dimension to different factors, which is described by a row of R. If one dimension is informative to only one ground truth factor, then this dimension is disentangled. This matches the mechanism of entropy. So we use the entropy to describe the disentanglement level of a dimension. We treat each row as a discrete distribution over factor index by normalization: $P_{ij} = R_{ij} / \sum_{k=0}^{K-1} R_{ik}$, where higher probability P_{ij} indicates that the dimension $\boldsymbol{c}_i$ encodes more information of the factor $\boldsymbol{v}_j$. Then the disentanglement score S_i for a latent dimension $\boldsymbol{c}_i$ is calculated as

$$S_i = 1 - H_K\left(P_{i\cdot}\right), \tag{2}$$

where $H_K\left(P_{i\cdot}\right) = -\sum_{k=0}^{K-1} P_{ik} \log_K P_{ik}$ is the entropy. S_i will be higher if $\boldsymbol{c}_i$ exhibits more informativeness to one factor while less relevance to other factors. Finally, to summarize the overall disentanglement of a representation model, MED score is the weighted average of the disentanglement scores for all dimensions as

$$\text{MED}\left(\boldsymbol{c}\right) = \sum_{i=0}^{D-1} \rho_i S_i \tag{3}$$

where $\rho_i = \sum_j R_{ij} / \sum_{ij} R_{ij}$ is the relative importance of each dimension.

Our proposed MED does not use any learnable classifier or regressor. Also, it inherits a DCI-style normalized importance matrix by taking all dimensions into calculation instead of using only one or two dimensions. These characteristics allow MED more robust to the latent dimensionality and eligible for a meaningful disentanglement measurement in high-dimensional space. Besides the advantages MED is designed to have, it is also very computationally efficient: for a 1000-d BYOL representation model on Cars3D, the evaluation with DCI with Gradient Boost Tree by DisLib takes more than 14 hours, while MED only takes less than 20 seconds on the same machine.

3.3 Partial Disentanglement Evaluation Metric

It is challenging to learn a fully disentangled representation by high-dimensional models without explicitly encouraging disentanglement, especially when there are fewer independent data factors than the number of latent dimensions. However, if a high-dimensional model has a subset of dimensions that disentangle well, it is still worth studying. This motivates us to design a version of MED to evaluate the partial disentanglement, which we name it as "Top-k MED".

The main difference is that we pick the most disentangled k dimensions for each factor and compute MED on this subset of representations. For each ground truth factor $\boldsymbol{v}_j$, $\mathcal{G}_j = \{i | \arg\max_m R_{im} = j\}$ is the set of latent dimensions emphasizing this factor. Then we pick the top k latent dimensions with the highest disentanglement score in each $\mathcal{G}_j$ to construct a subset

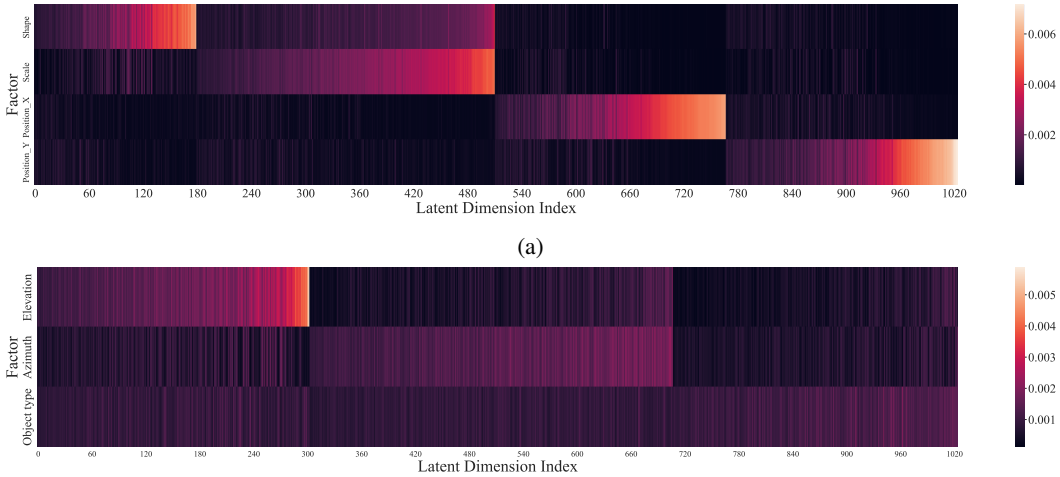

Figure 1: The mutual information heatmap between factors and the BYOL's latent dimensions on dSprites (a) and Cars3D (b). Each row is a ground truth factor and each column is a learned latent dimension. We normalize each row as we do in MED. Brighter straps denote higher MI. Latent dimensions are sorted for visualization purpose.

$\mathcal{P}_j = \{i | S_i \geq S^k,\ i \in \mathcal{G}_j\}$. Here, S_i is the disentanglement score of the latent dimension $\boldsymbol{c}_i$ defined in Equation 2. S^k is the kth highest disentanglement score in $\mathcal{G}_j$. Finally we obtain the subset $\mathcal{P} = \cup_j \mathcal{P}_j$. We take the sub-vector after selection $\tilde{\boldsymbol{c}} = \{\boldsymbol{c}_i\}_{i \in \mathcal{P}}$ as the representation to evaluate MED. The top-k MED is defined as MED$(\tilde{\boldsymbol{c}})$.

4 Understanding the learned representation

In this part, we qualitatively study the disentanglement of the learned representation by negative-free contrastive learning with BYOL as an example. Since contrastive methods are not generative models, it is hard to directly do factor-controlled pixel-wise reconstruction for visualization. Instead, we measure the mutual information between the learned representations and ground truth factors, which is also the foundation of our proposed MED metric.

4.1 Correspondence of Representation and Factor

To understand the disentanglement of a representation model, a basic question is how representation dimensions correspond to factors of data. After encoding an input image to a representation vector, we compute the normalized mutual information (MI) between each ground truth factor and each representation dimension, i.e. R in MED defined as in Equation 1, to measure the correspondence.

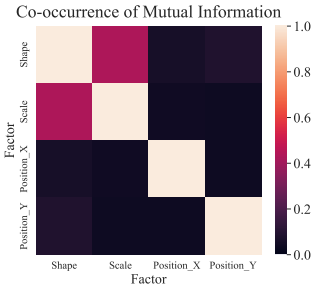

Figure 2: The visualization of normalized co-occurrence of mutual information on dSprites.

The mutual information between dimensions and the factors is included in Figure 1a and Figure 1b. The heatmap is actually the transpose of the importance matrix $R^\top$. As shown by the brightness of the entries, the informativeness of the latent dimensions varies greatly. In fact, we can identify three types of columns: columns with *single*, *multiple* and *no* bright elements, corresponding to three types of latent dimensions: *disentangled* dimensions, *entangled* dimensions and *uninformative* dimensions. The *disentangled* dimensions only capture information of one factor while the *entangled* dimensions encode multiple factors. In contrast, the *uninformative* dimensions can not represent factors independently. Note that the *entangled* and *uninformative* dimensions are non-negligible, hence the representations are not fully disentangled. However, by extracting the subset of *disentangled* dimensions we can derive a well-disentangled subspace.

First, we analyze the model on dSprites dataset. dSprites has five factors (shape, scale, orientation, position_x, and position_y). But the orientation is ill-defined with ambiguity. For example, it is

impossible to distinguish if a square rotates 0 degrees or 180 degrees. As shown in Figure 1a, the *disentangled* dimensions occupy a significant proportion, indicating an evident partially disentangled pattern. We note that the degree of disentanglement varies on different datasets. An example on Cars3D is shown in Figure 1b. Cars3D is a dataset with 183 different car objects rendered from 4 elevations and 24 azimuths and the ground truth factors are not fully independent of Cars3D (see Appendix B). It is extremely hard to represent its azimuth and object id with few dimensions. Thus its object-type row and azimuth row in Figure 1b are more spread out among multiple latent dimensions. This also shows the difficulty of understanding the disentanglement of the high-dimensional representation model on complicated datasets. Therefore, we will continue to conduct a quantitative evaluation with our proposed MED metric in Section 5. Such quantitative study on more datasets is available in Appendix B in the supplementary materials.

4.2 Uniqueness of Factor-Representation Correspondence

In the ideal pattern of disentanglement, a representation dimension should uniquely correspond to only one factor. Now we show to what extent multiple factors are responded to by a single representation dimension in a well-disentangled subspace. Given the mutual information between the representation and the ith factor, noted as I_i, i.e., the ith row in the top-k version of Figure 1a, a good indicator of the uniqueness of factor-representation correspondence is the normalized co-occurrence of mutual information between the i_1^{th} factor and the i_2^{th} factor, which is defined as

$$\widehat{C}_{i_1,i_2} = \frac{\langle I_{i_1}, I_{i_2} \rangle}{||I_{i_1}||_2 \cdot ||I_{i_2}||_2} = \frac{\sum_{d=0}^{K \cdot k-1} I(\tilde{\boldsymbol{c}}_d, \boldsymbol{v}_{i_1}) I(\tilde{\boldsymbol{c}}_d, \boldsymbol{v}_{i_2})}{||I_{i_1}||_2 \cdot ||I_{i_2}||_2}, \tag{4}$$

where $\tilde{\boldsymbol{c}}$ is the representation after the selection process in Section 3.3. We visualize the normalized co-occurrence of mutual information among the four factors by the learned representation in Figure 2. It agrees that a dimension usually encodes only one factor. Moreover, it indicates that the learned representation tends to encode shape and scale together, which also agrees with the intuitive analysis of the independence of factor pairs. For example, the shape and scale of dSprites objects are not disentangled and independent because objects with the same scale value but in different shapes have different pixel area.

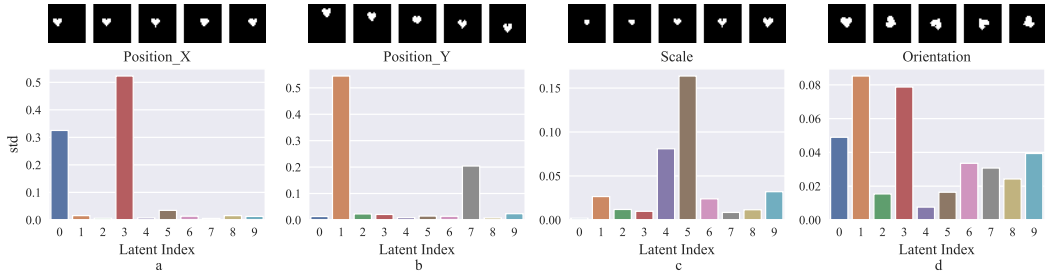

Figure 3: Representation variation when manipulating one factor only in the dimension-reduced version. In (a), (b), (c), *position_x*, *position_y* and *scale* are manipulated respectively and only cause one dimension significantly variate. While, in (d), when manipulating the ill-defined factor *orientation*, multiple dimensions variate.

4.3 Influence by Manipulating Factors

Another intuitive perspective to studying the relationship between representation and factors is the influence on representation when manipulating the factors. Given that the original representation vector dimension is much higher than the number of factors, we first make the representation more compact to have a more concise illustration. Here, we reduce the representation dimension by the selection process of top-k MED described in Section 3.3.

Figure 3 shows the result of representation vector variation when changing only one factor at once. We take BYOL on dSprites as a representative. Here we set $k = 2$ for top-k MED, i.e., we pick the 2 most disentangled dimensions for each factor and derive a 10-dim representation. Given images with one factor traversing through all its support values while fixing the other factors, we generate the 10-dim representation vectors from them. Then, we compute the variance of each of the 10 dimensions across the images, leading to 10 scalars. The larger the variance is, the more

that dimension responds to the factor. Figure 3(a), (b) and (c) show how the reduced representation vector changes when manipulating *position_x*, *position_y*, *scale* factor respectively. Note that we set $k = 2$, therefore it shows good disentanglement that exactly two representation dimensions have high variation. However, in Figure 3(d) we show a failure mode of the ill-defined factor *orientation* that change of factor causes multiple dimensions of reduced representation to have large variations, indicating that this factor is represented in an entangled way. From the results, we observe that manipulating different well-defined independent factors causes evident variance on disjoint sets of dimensions. And it demonstrates the existence of a well-disentangled subset of latent dimensions.

We also conduct a similar qualitative study where the representation dimension is reduced by the unsupervised PCA technique and that shows the similar pattern observed here. The details are provided in Section B.4 of the Appendix in the supplementary materials.

5 Quantitative Evaluation

In this section, we conduct a quantitative evaluation to study the disentanglement property of contrastive learning methods. We first introduce the experiment setup in Section 5.1. Then we show quantitative results under both MED and existing disentanglement metrics in Section 5.2 which shows the disagreement of existing metrics to support the necessity of proposing MED. Finally, we make a full quantitative benchmark of methods of interest with MED in Section 5.3 and ablation study about the dimension of the representation model in Section 5.4.

5.1 Experiments Setup

The details for reproducibility are introduced in Appendix Section A included in the supplementary materials. Here we provide a brief description of the setup of the experiments.

Datasets. Representation disentanglement is usually evaluated on synthetic datasets, such as dSprites (36), Cars3D (39), Shapes3D (6), and SmallNORB (29). Besides those datasets, we also include a real-world dataset CelebA (32). CelebA contains human face images with 40 binary attributes. The attributes include fine-grained properties of the human face, such as whether wearing glasses or having wavy hair. We include the details of dataset factors in the supplementary materials.

Evaluation Protocol. We conduct experiments with both MED and the existing metrics to reveal the disagreement between existing metrics. Then we use MED as the main metric to study the disentanglement of contrastive learning methods. The implementation of evaluation metrics is adapted from the protocol provided by DisLib (34). All results are calculated with three random seeds and we report both the average score and the standard deviation. More details are introduced in Appendix Section A.2 in the supplementary materials.

Reference Methods. We investigate most of the popular disentangled representation learning methods as studied in the standard benchmark of DisLib (34). Besides, we also compare with a recently proposed ICA method called ICE-BeeM (22). Since we do not assume the ground truth factors are known during the training, we use its unconditional version. We term it EBM (energy-based model). For the contrastive learning methods, we take not just negative-free methods, such as BYOL (40), Barlow Twins (51) and SimSiam (8), into the benchmark, but also those using negative samples, such as MoCo and MoCov2 (15).

Model Implementation. All methods use a shared architecture of encoder network as explained in the appendix in supplementary materials. The latent dimension of contrastive learning methods is set to 1000 since they require a high-dimensional latent space to work. For other methods, the latent dimension is set to be 10 on synthetic datasets as in DisLib and 128 on CelebA dataset when evaluating with MED. For the evaluation with Top-k MED, the dimension of all methods is set to be 1000-d for fairness. On dSprites, Cars3D and SmallNORB, we acquire checkpoints from DisLib if they are provided. We train our checkpoints on CelebA and Shapes3D.

5.2 Disagreement of Existing Metrics

As there is not a uniform definition of the "disentanglement" concept yet, existing disentanglement metrics are motivated by different desired properties of a disentangled representation. We rely on them before because they show good agreement as proven by the large-scale experiments in DisLib (34). However, when we extend the disentanglement study beyond low-dimensional scenarios, this agreement does not stand anymore. We select representative low-dimensional VAE-based methods and a high-dimensional BYOL model to evaluate on a representative dataset, the SmallNORB

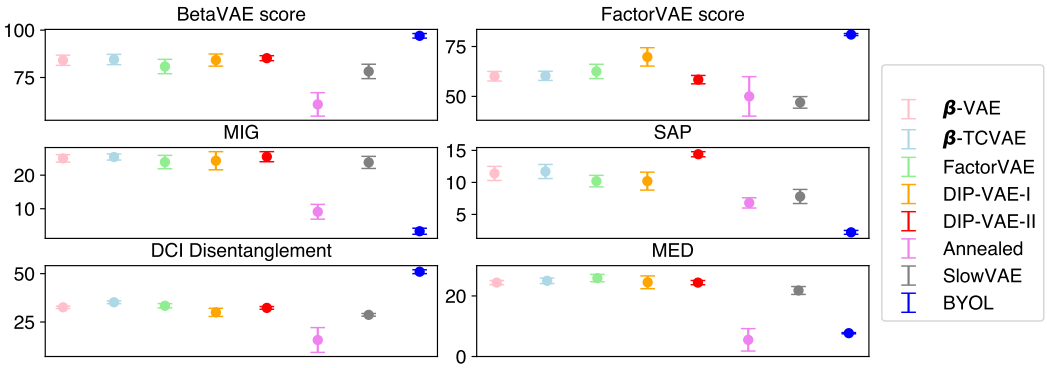

Figure 4: Evaluation under multiple metrics on SmallNORB. It shows clear disagreement of existing metrics when taking high-dimensional BYOL method into consideration. Y-axis is the corresponding disentanglement scores.

dataset. The results are shown in Figure 4. The results show a significant disagreement among metrics on the BYOL method, while they agree on the low-dimensional VAE methods. Aligned with our analysis in Section 3.2, BetaVAE score, FactorVAE score, and DCI overestimate the disentanglement degree of the high-dimensional model while MIG and SAP underestimate. This explains why we need a new metric if we want to evaluate the disentanglement of contrastive learning. Since previous metrics fail to evaluate high-dimensional latent space, we opt to use MED as the main evaluation metric in the following sections.

Table 1: MED and Top-k MED scores on multiple datasets. Methods in gray are contrastive self-supervised learning methods. The latent dimension of VAE methods with * is 1000.

Metrics	Model	dSprites	Shapes3D	Cars3D	SmallNORB	CelebA
MED	β-VAE	32.6 (10.0)	12.9 (3.5)	29.0 (2.2)	24.4 (0.7)	3.3 (0.5)
	β-TCVAE	31.8 (7.4)	**13.7 (0.9)**	**33.0 (3.8)**	25.0 (0.9)	4.7 (0.1)
	FactorVAE	32.5 (10.1)	0.7 (0.9)	29.1 (3.0)	**25.9 (1.2)**	0.6 (0.6)
	DIP-VAE-I	18.8 (5.6)	10.3 (0.9)	19.4 (3.3)	24.5 (2.1)	3.7 (0.2)
	DIP-VAE-II	14.7 (5.5)	–	16.7 (4.1)	24.4 (0.6)	–
	AnnealedVAE	**35.8 (0.8)**	–	15.5 (2.5)	5.5 (3.7)	–
	EBM	6.8 (4.0)	2.1 (2.6)	–	2.3 (1.7)	–
	MoCo	4.2 (0.5)	6.1 (0.1)	8.6 (0.4)	4.9 (0.1)	**5.8 (0.1)**
	MoCov2	3.5 (1.4)	4.2 (0.4)	6.5 (0.4)	3.3 (0.2)	4.8 (0.2)
	BarlowTwins	6.0 (0.3)	6.4 (0.3)	5.6 (1.4)	6.1 (0.1)	4.2 (0.2)
	SimSiam	26.5 (0.1)	12.3 (0.9)	10.4 (0.1)	10.7 (1.1)	5.3 (0.4)
	BYOL	31.3 (0.4)	6.0 (0.5)	9.7 (0.5)	7.7 (0.2)	4.8 (0.4)
Top-k MED	β-VAE*	16.6 (6.2)	19.2 (1.4)	29.2 (2.0)	15.8 (2.1)	4.5 (0.3)
	β-TCVAE*	11.2 (0.4)	25.0 (0.5)	20.0 (1.8)	23.0 (0.6)	3.6 (0.2)
	FactorVAE*	3.2 (3.8)	8.2 (4.0)	7.8 (1.6)	4.8 (1.0)	5.0 (0.3)
	DIP-VAE-I*	7.0 (1.3)	16.2 (0.9)	24.6 (2.2)	20.9 (2.7)	2.5 (0.9)
	MoCo	16.1 (2.0)	18.1 (0.6)	26.6 (1.6)	17.9 (0.8)	**7.9 (0.1)**
	MoCov2	14.7 (1.0)	13.6 (1.7)	24.5 (2.1)	15.1 (0.9)	6.6 (0.7)
	BarlowTwins	21.7 (1.3)	20.0 (0.3)	23.8 (2.5)	24.5 (1.5)	5.7 (0.2)
	SimSiam	39.1 (0.4)	**30.0 (2.0)**	**32.7 (2.3)**	**28.4 (1.9)**	7.2 (0.6)
	BYOL	**53.7 (0.7)**	19.7 (1.3)	31.8 (1.3)	25.7 (0.3)	6.8 (0.7)

5.3 Disentanglement Benchmark with Contrastive Learning Methods

For the benchmarking of disentanglement, we use both MED and the partial version of MED, i.e. Top-k MED. For Top-k MED, we set the MED partial evaluation hyperparameter $k = 2$ for dSprites, Shape3D, and SmallNORB, and $k = 3$ for Cars3D and CelebA. The values of k are chosen such that the selected dimensions are roughly close to the latent space dimension of the low-dimensional reference methods. And we extend the latent dimension of all methods to 1000 when evaluating Top-k MED for fairness. We note that the hyperparameters we tried fail to train good EBM weights on Cars3D and CelebA, so we keep that empty. The results are in Table 1.

In the upper part of Table 1, we show the MED score for previous disentangled methods as well as contrastive methods. We find that contrastive methods achieve lower disentanglement scores on 3 of the 5 datasets (Shapes3D, Cars3D, SmallNORB). Contrastive methods achieve slightly higher disentanglement scores on CelebA. On dSprites, some negative-free contrastive methods (SimSiam, BYOL) achieve scores close to SOTA, but the other contrastive methods' score is much lower. In summary, in most cases, contrastive methods achieve inferior disentanglement property than the best methods; in few settings, it achieves scores comparable to SOTA scores.

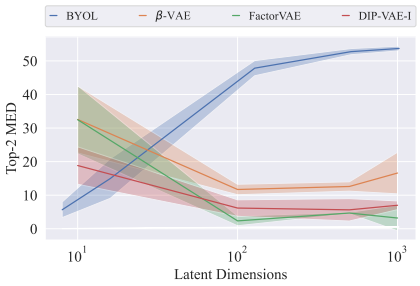

Figure 5: The influence of representation dimension to the top-2 MED score on dSprites dataset. The disentanglement is enhanced with increasing representation dimension for BYOL method but decreased for all tested VAE-based methods. The minimum dimension is 10 for VAE-based methods, and 8 for BYOL.

This result seems to be disappointing, but it is not surprising, since contrastive methods are not explicitly designed to maximize the feature disentanglement. Further, since the underlying number of data factors is usually quite small, on the order of 10, the 1000 dimensional feature space will likely have dimensions that are not related to the ground truth factors, or a combination of the ground truth factors. This result does not contradict with Zimmermann et al. (53) since (1) their conclusion is the learned feature is a linear transformation of the ground truth factors, and that feature doesn't necessarily disentangle (2) they use augmentations on factors that can not be done in practice.

The lower part of Table 1 shows top-k MED measurements on various methods. We find that contrastive learning methods (especially the negative-free ones) in general show a better disentanglement in a selected subspace and the disentanglement is stronger than the reference methods. This shows that there exists a subspace in the learned representation that is well disentangled. Moreover, when we compare the subspace in contrastive methods (gray part in the lower section of Table 5.3) to the traditional approach that directly trains a low-dimensional latent space (non-gray part in the upper section of Table 5.3), we find that the disentanglement of the former is usually better than or on par with that of the latter. This means that we probably should not constraint the dimension of the latent space and require it to be fully disentangled, but rather should encourage to use high dimensional latent spaces and only require it to have a subset with good disentangle properties.

To conclude, we find that contrastive methods, especially the negative-free ones, do not have a fully disentangled feature space, but there exists a subspace that is well disentangled, sometimes much better than the previous SOTA approaches.

5.4 Influence of Dimension

As representation dimension is found as a core variable in disentangled representation learning, we evaluate the influence of dimension over disentanglement by top-2 MED on dSprites with BYOL as an example versus VAE-based methods. The results are shown in Figure 5. We find that the BYOL's top-2 MED score increases as the latent dimension increases and plateaus at 512 dimensions. This is consistent with previous literature on the difficulty to train contrastive models with low latent dimensions (14) which shows lower latent dimension leads to inferior linear accuracy. We further show that a lower latent dimension leads to a less disentangled subspace as well. On the contrary, we also note that the VAE methods fail to scale to higher latent dimensions. We have spent a reasonable effort to tune VAEs' hyperparameters, but we still find a large gap between higher-dimensional VAEs and their 10-dim version. This suggests the gap between existing disentangled representation methods and the real-world data complexity, which can not be represented in a limited dimension anymore.

6 Conclusion

In this paper, we provide an empirical study of the disentanglement property of contrastive learning without negatives for the first time. In the high-dimensional space, we find the difficulty of adopting the existing disentanglement metrics. Therefore, we propose a new metric MED and top-k MED to evaluate disentanglement based on mutual information. The evaluation shows that even without negative samples, contrastive learning can learn a well-disentangled subset of representation. Recently, the study of contrastive learning, or general self-supervised learning, is still motivated by empirical observations. We hope our work can reveal some clues to motivate future theoretical justifications.

References

[1] A. Achille, T. Eccles, L. Matthey, C. P. Burgess, N. Watters, A. Lerchner, and I. Higgins. Life-long disentangled representation learning with cross-domain latent homologies. *arXiv preprint arXiv:1808.06508*, 2018.

[2] S. Arora, H. Khandeparkar, M. Khodak, O. Plevrakis, and N. Saunshi. A theoretical analysis of contrastive unsupervised representation learning. *arXiv preprint arXiv:1902.09229*, 2019.

[3] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6541–6549, 2017.

[4] Y. Bengio, Y. LeCun, et al. Scaling learning algorithms towards ai. *Large-scale kernel machines*, 34(5):1–41, 2007.

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

[6] C. Burgess and H. Kim. 3d shapes dataset. https://github.com/deepmind/3dshapes-dataset/, 2018.

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

[8] X. Chen and K. He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021.

[9] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 2180–2188, 2016.

[10] A. Dosovitskiy, J. T. Springenberg, M. Riedmiller, and T. Brox. Discriminative unsupervised feature learning with convolutional neural networks. *Advances in neural information processing systems*, 27:766–774, 2014.

[11] C. Eastwood and C. K. Williams. A framework for the quantitative evaluation of disentangled representations. In *ICML*, 2018.

[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[13] I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT press, 2016.

[14] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020.

[15] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

[16] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. *arXiv preprint*, 2016.

[17] A. Hyvarinen and H. Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. *Advances in Neural Information Processing Systems*, 29:3765–3773, 2016.

[18] A. Hyvärinen and E. Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430, 2000.

[19] A. Hyvarinen, H. Sasaki, and R. Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 859–868. PMLR, 2019.

[20] T. Jakab, A. Gupta, H. Bilen, and A. Vedaldi. Unsupervised learning of object landmarks through conditional image generation. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 4020–4031, 2018.

[21] I. Khemakhem, D. Kingma, R. Monti, and A. Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217. PMLR, 2020.

[22] I. Khemakhem, R. P. Monti, D. P. Kingma, and A. Hyvärinen. Ice-beem: Identifiable conditional energy-based deep models based on nonlinear ica. *arXiv preprint arXiv:2002.11537*, 2020.

[23] H. Kim and A. Mnih. Disentangling by factorising. In *International Conference on Machine Learning*, pages 2649–2658. PMLR, 2018.

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

[25] D. Klindt, L. Schott, Y. Sharma, I. Ustyuzhaninov, W. Brendel, M. Bethge, and D. Paiton. Towards nonlinear disentanglement in natural data with temporal sparse coding. *arXiv preprint arXiv:2007.10930*, 2020.

[26] T. D. Kulkarni, W. Whitney, P. Kohli, and J. B. Tenenbaum. Deep convolutional inverse graphics network. *arXiv preprint arXiv:1503.03167*, 2015.

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

[28] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40, 2017.

[29] Y. LeCun, F. J. Huang, and L. Bottou. Learning methods for generic object recognition with invariance to pose and lighting. *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2:II–104 Vol.2, 2004.

[30] J. D. Lee, Q. Lei, N. Saunshi, and J. Zhuo. Predicting what you already know helps: Provable self-supervised learning. *arXiv preprint arXiv:2008.01064*, 2020.

[31] Z. Lin, K. Thekumparampil, G. Fanti, and S. Oh. Infogan-cr and modelcentrality: Self-supervised model training and selection for disentangling gans. In *International Conference on Machine Learning*, pages 6127–6139. PMLR, 2020.

[32] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.

[33] Z. Liu, P. Luo, X. Wang, and X. Tang. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15(2018):11, 2018.

[34] F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, and O. Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *ICML*, 2019.

[35] F. Locatello, B. Poole, G. Rätsch, B. Schölkopf, O. Bachem, and M. Tschannen. Weakly-supervised disentanglement without compromises. In *International Conference on Machine Learning*, pages 6348–6359. PMLR, 2020.

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

[37] J. Peters, D. Janzing, and B. Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

[38] S. Purushwalkam and A. Gupta. Demystifying contrastive self-supervised learning: Invariances, augmentations and dataset biases. *arXiv preprint arXiv:2007.13916*, 2020.

[39] S. E. Reed, Y. Zhang, Y. Zhang, and H. Lee. Deep visual analogy-making. *Advances in neural information processing systems*, 28:1252–1260, 2015.

[40] P. H. Richemond, J.-B. Grill, F. Altché, C. Tallec, F. Strub, A. Brock, S. Smith, S. De, R. Pascanu, B. Piot, et al. Byol works even without batch statistics. *arXiv preprint arXiv:2010.10241*, 2020.

[41] K. Ridgeway and M. C. Mozer. Learning deep disentangled embeddings with the f-statistic loss. *arXiv preprint arXiv:1802.05312*, 2018.

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

[43] Y. Tian, C. Sun, B. Poole, D. Krishnan, C. Schmid, and P. Isola. What makes for good views for contrastive learning? *arXiv preprint arXiv:2005.10243*, 2020.

[44] C. Tosh, A. Krishnamurthy, and D. Hsu. Contrastive learning, multi-view redundancy, and linear models. In *Algorithmic Learning Theory*, pages 1179–1206. PMLR, 2021.

[45] Y.-H. H. Tsai, Y. Wu, R. Salakhutdinov, and L.-P. Morency. Demystifying self-supervised learning: An information-theoretical framework. *arXiv e-prints*, pages arXiv–2006, 2020.

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

[47] D. Ulyanov, A. Vedaldi, and V. Lempitsky. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6924–6932, 2017.

[48] S. van Steenkiste, F. Locatello, J. Schmidhuber, and O. Bachem. Are disentangled representations helpful for abstract visual reasoning? *arXiv:1905.12506*, 2019.

[49] T. Wang and P. Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR, 2020.

[50] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018.

[51] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021.

[52] N. Zhao, Z. Wu, R. W. Lau, and S. Lin. What makes instance discrimination good for transfer learning? *arXiv preprint arXiv:2006.06606*, 2020.

[53] R. S. Zimmermann, Y. Sharma, S. Schneider, M. Bethge, and W. Brendel. Contrastive learning inverts the data generating process. *arXiv:2102.08850*, 2021.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?
   [Yes] The three main contributions listed at the end of the introduction reflect the paper's contributions comprehensively.
   (b) Did you describe the limitations of your work?
   [Yes] The limitations are discussed in the Section of **Limitation** in the supplementary materials.
   (c) Did you discuss any potential negative societal impacts of your work?
   [No] We don't recognize obvious negative societal impacts of our work.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them?
   [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results?
   [Yes] The only theoretical part is related to the definition of our proposed MED metric, we have stated the definition of involved variables and the assumption, which is aligned to the default settings of DisLib (34).
   (b) Did you include complete proofs of all theoretical results?
   [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)?
   [Yes] All required codebases, hyperparameters, and other instructions are provided in the Section of **Reproducibility** in the Appendix in supplementary materials.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)?
   [Yes] As all datasets we use are publicly available, we follow the formal version of them, including the splits. The data preprocessing is provided in the Section of **Reproducibility** or the codebases we provide in the supplementary materials.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)?
   [Yes] We ran experiments with multiple random seeds and provide the standard deviation in every result entry from our experiments.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)?
   [Yes] The details are provided in the Section of **Reproducibility** in Appendix included in the supplementary materials.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators?
   [Yes] We cite the related works and also provide the links to open-sourced code we use.
   (b) Did you mention the license of the assets?
   [Yes] We don't explicitly provide the licenses of all in the main text due to the page limit. But we have confirmed that the license is available in the links to the code we provide if they have chosen one.
   (c) Did you include any new assets either in the supplemental material or as a URL?
   [Yes] We include the links to open-sourced code we use in this work.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating?
   [No] We mostly work on public synthetic datasets. For the only real-world dataset CelebA (33), the resources of data and the applied agreement of data use are available at their webpage [1].

[1] https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content?
[No] We only use public datasets and can hardly provide full report about this. The related information in the only real-world dataset CelebA is available as mentioned above.

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable?
[N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable?
[N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation?
[N/A]