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 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; 39; 13; 4; 44; 30; 49). A disentangled representation matches how humans understand the world and allows us to use much fewer labels to learn challenging downstream tasks (51). Previous disentangled learning research is usually limited to using generative models (26; 16; 25; 12; 9; 33), on simple synthetic datasets (38; 6) and with 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 (52; 56) 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. And no contrastive learning methods have been evaluated in the standard disentanglement benchmarks yet.

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; 25; 7; 11; 29) are demonstrated well aligned in low-dimensional space (36), 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 (35). 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; 28), to make the downstream tasks learned more easily (51) 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 (19) usually assumes that the pattern of noise (18; 23) or some additional auxiliary variables (20; 24) 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; 25; 7; 29) and Generative Adversarial Networks (GAN) (12; 9). These two lines of study focus on different but relevant (27) 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 (36) (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 (29), MIG (7), BetaVAE score (16), FactorVAE score (25) and Modularity (43). DisLib finds that the five metrics except for Modularity have good agreement in evaluating disentanglement quality. Till now, no widely accepted definition of "disentanglement" is available. So the empirical agreement of metrics and their shared high-level intention of disentanglement, i.e. the degree to which a representation dimension only responds to a single factor, make the basis of quantitatively evaluating disentanglement. 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 thus is a proper tool to evaluate the disentanglement of contrastive learning methods.

Contrastive Learning. Contrastive learning (CL) 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 (53). Recently, some works try to understand contrastive learning either theoretically (52; 2; 48; 47; 32) or empirically (45; 55; 40). Zimmermann et al. suggests that the contrastive method inverts a data generation process, which is related to the disentanglement property. This conclusion is based on 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). There is still no work connecting CL methods with standard disentanglement benchmarks. On the other hand, negative-free CL (42; 8; 54) has its own advantages, such as avoiding model collapse (46; 17; 22), but its disentanglement property remains unexplored, either empirically or theoretically. Thus, we aim to empirically study whether negative-free CL can learn 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) 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 and negative samples to encourage discriminative representations. They 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 of the same architecture. However, recent works on this line, such as BYOL (42), SimSiam (8) and Barlow Twins (54), have their own unique designs. The development of CL still heavily relies on empirical observation to learn non-trivial representations.

As the first and the most classic method in this line, BYOL (42) 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 (54) 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 (52). 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 as in DisLib. But such an empirical agreement is the basis of this community to make quantitative evaluations without a uniformly shared definition of "disentanglement". So 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 (MI) between latent dimensions and dataset ground truth factors as MI is a widely accepted tool to measure correlation of variables and is not biased to models of different dimensions. 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 overall informativeness to all factors. This operation fixes the gap between models of different dimensions by estimating the relative importance of a single dimension.

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 perfectly 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".

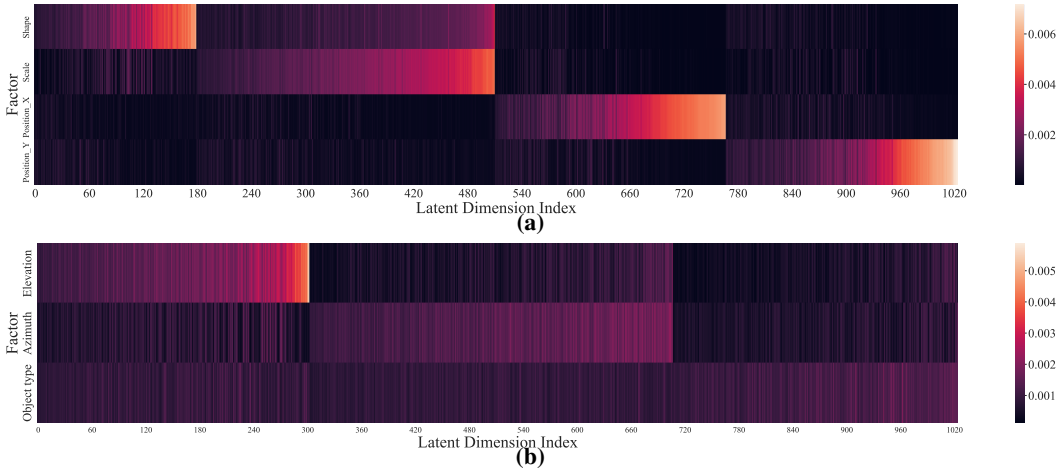


Figure 1: The mutual information heatmap between factors and the BYOL's latent dimensions on dSprites (a) and Cars3D (b). Each row is a ground truth factor and each column is a learned latent dimension. We normalize each row as we do in MED. Brighter straps denote higher MI. Latent dimensions are sorted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

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 $\mathcal{P}_j = \left\{i | S_i \geq S^k,\ i \in \mathcal{G}_j\right\}$. 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. More qualitative study is available in Appendix B.

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.

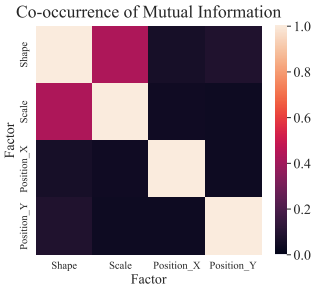


Figure 2: The visualization of normalized co-occurrence of mutual information on dSprites.

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.

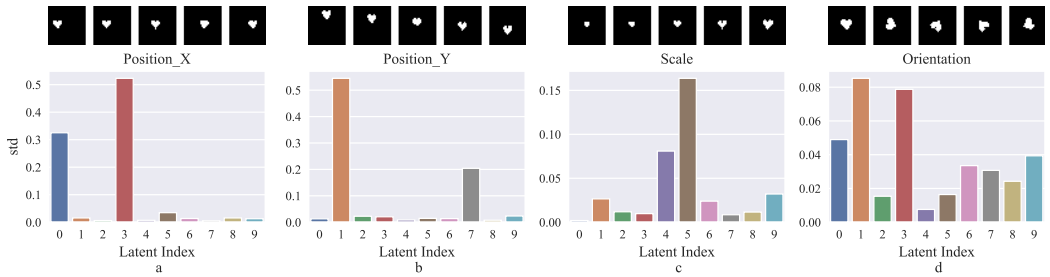


Figure 3: Representation variation when manipulating one factor only in the dimension-reduced version. In (a), (b), (c), *position_x*, *position_y* and *scale* are manipulated respectively and only cause one dimension significantly variate. While, in (d), when manipulating the ill-defined factor *orientation*, 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 quantitative studies on 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. More quantitative study is available in Appendix C.

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 (38), Cars3D (41), Shapes3D (6), and SmallNORB (31). Besides those datasets, we also include a real-world dataset CelebA (34). 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 (36). 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 (36). Besides, we also compare with a recently proposed ICA method called ICE-BeeM (24). 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 (42), Barlow Twins (54) 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 (36). 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 dataset. The results are in Figure 4, which show a significant disagreement among metrics on BYOL, while they agree on the low-dimensional VAE methods. Aligned with our analysis in Section 3.2,

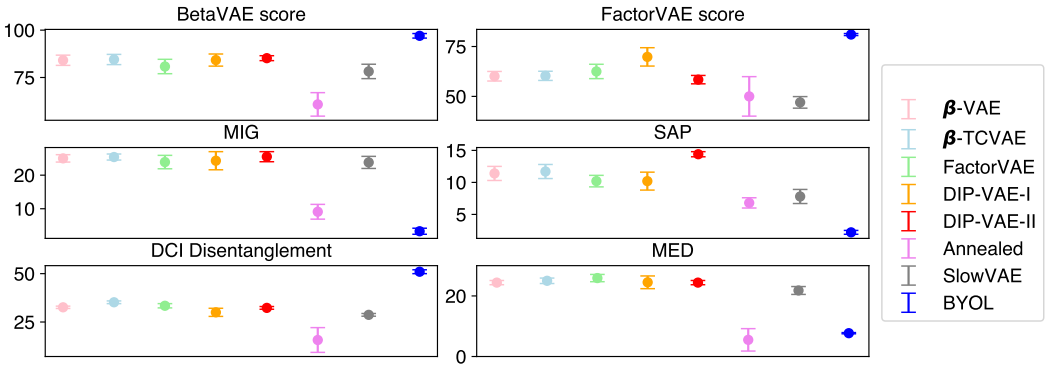


Figure 4: Evaluation under multiple metrics on SmallNORB. It shows clear disagreement of existing metrics when taking high-dimensional BYOL method into consideration. Y-axis is the corresponding disentanglement scores.

BetaVAE score, FactorVAE score, and DCI overestimate the disentanglement degree of the high-dimensional model while MIG and SAP underestimate. Since previous metrics fail to evaluate high-dimensional models, we opt to use MED as the main evaluation metric in the following sections. Please refer to Appendix C and G for more evidence of the disagreement of existing metrics.

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. We also encourage to read the results together with existing metrics in Table 4 in Appendix C.

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.

This result seems disappointing, but not surprising, since contrastive methods are not explicitly designed to maximize the feature disentanglement. Further, since the underlying number of factors is usually quite small, on the order of 10, the 1000-d 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. (56) since (1) their conclusion is the learned feature is a linear transformation of the ground truth factors, which doesn't necessarily disentangle (2) they use augmentations on factors that can not be done in practice.

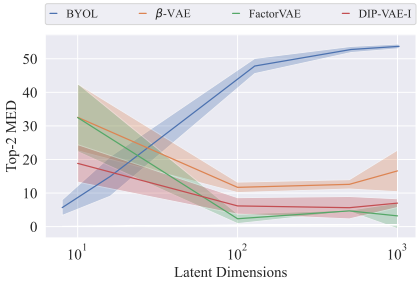


Figure 5: The influence of representation dimension to the top-2 MED score on dSprites dataset. The disentanglement is enhanced with increasing representation dimension for BYOL method but decreased for all tested VAE-based methods. The minimum dimension is 10 for VAE-based methods, and 8 for BYOL.

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, which may serve as a proxy between CL, which requires high dimension to train, and a more compact low-dimensional model from its subspace.

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 also study its influence on the full space regarding MED in Figure 12 in Appendix C. 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] T. Hua, W. Wang, Z. Xue, S. Ren, Y. Wang, and H. Zhao. On feature decorrelation in self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9598–9608, 2021.

[18] 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.

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

[20] 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.

[21] 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.

[22] L. Jing, P. Vincent, Y. LeCun, and Y. Tian. Understanding dimensional collapse in contrastive self-supervised learning. *arXiv preprint arXiv:2110.09348*, 2021.

[23] 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.

[24] 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.

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

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

[27] 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.

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

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

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

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

[36] 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.

[37] 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.

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

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

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

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

[42] 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.

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

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

[45] 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.

[46] Y. Tian, L. Yu, X. Chen, and S. Ganguli. Understanding self-supervised learning with dual deep networks. *arXiv preprint arXiv:2010.00578*, 2020.

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

[48] 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.

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

[50] 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.

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

[52] 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.

[53] 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.

[54] 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.

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

[56] 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 (36).
   (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 (35), 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]

A Reproducibility

In this section, we provide the information required to reproduce our results reported in the main text. And we commit to making the code implementation and evaluating checkpoints public. Our experiments are run on a machine with AMD Ryzen Threadripper 3970X 32-Core Processor and GeForce RTX 3090 GPU.

Contrastive Learning implementation For the implementation details of contrastive learning, please refer to Appendix A.1. The model architecture, training setups, and dataset preprocessing are all explained in detail. Our implementations are based some public and official implementations of MoCo/MoCov2 [2], BYOL/ SimSiam [3] and Barlow Twins [4].

VAE methods implementation For evaluation on synthetic datasets, i.e., dSprites, Cars3D, Small-NORB, and Shapes3D, the disentanglement score is from the original logs of DisLib Locatello et al. [5]. In the released logs, each method has different training configurations, and our reported result is from the configuration with the highest average performance overall provided random seeds. For evaluation on CelebA dataset, we follow an open-sourced implementation in Pytorch [6] and align the encoder architecture of all methods to be the same as described in Appendix A.1. For the results on Shapes3D, because DisLib does not release the pretrained checkpoints, we use the same open-sourced implementation to reproduce with the configuration indicated by DisLib. Parameters are kept as the default well-tuned version in the provided implementation. When the latent dimension is 1000, training of BetaTC VAE will collapse with the default hyperparameters, we have to decrease the β to 3.0 to work it around.

GAN methods implementation Limited by the text length, we do not include the performance of GAN methods in the main text, but we will report some in the following appendix content. It is hard to include GAN methods' performance in the benchmark as the training is not always stable and the discriminator weights are usually not provided in many public codebases. When evaluating on synthetic datasets, the FactorVAE scores of InforGAN, IB-GAN, and InfoGAN-CR are provided in the paper of Lin et al.. But the evaluation of other metrics in Lin et al. uses a not aligned settings with Locatello et al., so we check its officially release [7] to reevaluate the provided implementation and model weights under the unified evaluation setup. We perform the same evaluation process for results on the CelebA dataset.

Energy-based Model (EBM) We refer to the implementation of ICE-BeeM (24) for this method. We use the officially released codebase for it [8]. The encoder implementation has been aligned with our default already. The only modification we make is to use the unconditional version instead of its default conditional version in loss computation to satisfy the fully unsupervised settings. Please refer to the **runners/real_data_runner.py** file of the codebase for details.

Evaluation Protocol For MED, we first compute MI following the implementation of MIG by DisLib (36). Then we calculate the entropy disentanglement score in the same way as the DCI Disentanglement score in DisLib. For other disentanglement metrics evaluation, we use the implementation of DisLib. The settings of some important parameters rather than our proposed MED are provided in Appendix A.2.

A.1 Implementation of contrastive learning model

Architecture To make a fair comparison with previous methods, we follow the encoder architecture in Factor VAE (25). The pipeline details are shown in Table 2. After each convolutional layer in the figure, there is a ReLU activation layer and a group normalization (group number = 4) layer for BYOL. So, the encoder is a stack of (Conv-ReLU-GN) blocks. For other contrastive learning methods, we keep the default batch normalization to replace GN. By default, the final output channel

[2] https://github.com/facebookresearch/moco
[3] https://github.com/lucidrains/byol-pytorch
[4] https://github.com/facebookresearch/barlowtwins
[5] https://github.com/google-research/disentanglement_lib
[6] https://github.com/AntixK/PyTorch-VAE
[7] https://github.com/fjxmlzn/InfoGAN-CR
[8] https://github.com/ilkhem/icebeem

number is 1000, i.e, $D = 1000$. For other details of contrastive learning methods, we follow the convention in their official implementations.

Besides the representation network (encoder), BYOL also has a projector network and a predictor network. Both of them consist of a pipeline "Linear $\longrightarrow$ BN $\longrightarrow$ ReLU $\longrightarrow$ Linear". The projection dimension is 256, and the hidden dimension of the projector is 4096. The predictor keeps a 256-dimensional feature vector in its pipeline.

Table 2: The encoder architecture for our implemented contrastive learning methods on synthetic datasets. Besides, there is a ReLU activation layer and a possible normalization layer following each convolutional layer to create a stack of (Conv-ReLU-Norm) blocks.

Encoder
input: 64×64 images
pipeline:
4×4 conv, stride 2, 32-channel
4×4 conv, stride 2, 32-channel
4×4 conv, stride 2, 64-channel
4×4 conv, stride 2, 64-channel
4×4 conv, stride 2, 128-channel
1×1 conv, stride 1, D-channel

Training settings We make minor modifications to the training setting of default BYOL to apply to contrastive learning methods without negative samples. For training on all datasets, the images are resized to 64x64. For data preprocessing, we copy 1-channel images of dSprites and SmallNORB to 3-channel. During the training stage, we use such a pipeline of augmentation (in *PyTorch*-style):

1. *RandomApply(transforms.ColorJitter(0.8, 0.8, 0.8, 0.2), p=0.3)*
2. *RandomHorizontalFlip()*
3. *RandomApply(transforms.GaussianBlur((3,3), (1.0, 2.0)), p=0.2)*
4. *RandomResizeCrop(size=(64, 64), scale=(0.6,1.0))*
5. normalization.

For the normalization, the pixel value of images from dSprites and SmallNORB is uniformly normalized from [0,255] to [0,1.0]. For Cars3D, Shapes3D, and CelebA, we adopt the commonly used Imagenet-statistic normalization for preprocessing the image values.

During training, we use Adam optimizer by default, whose learning rate is $3e-4$ without weight decay. The batch size is set to be 512 without exceptional notation. For evaluation on dSprites, Shapes3D, and CelebA, we select the weights after training for 15 epochs for evaluation. We select the weights after training for 140 epochs for evaluation on Cars3D and the weights of the 200th epoch on SmallNORB considering the small scale of these two datasets.

To decrease the influence of randomness, we train each model configuration multiple times with different random seeds (seed=0, 1, 2). We report the average and standard deviation. To be precise, as our implementation is based on Pytorch, we initialize the libraries of *numpy*, *torch*, *torch.cuda*, and *random* with the same random seeds.

A.2 Evaluation Metrics

In the main text, we compare the evaluation metrics provided in the DisLib protocol with our proposed MED metric. Here we provide more details about them. Moreover, we would conduct evaluations under all of them in the next section.

BetaVAE Metrics Introduced in Higgins et al. (16), BetaVAE score assumes each dimension corresponds to one category in a linear classifier. Representations are obtained after the generated samples with only one factor fixed. Calculating the summation of the divergence between different representations and putting this result into a linear classifier, we train a model that possibly outputs the corresponding k. The accuracy of this linear model is the value of BetaVAE metric.

Table 3: The factors on all the datasets we investigate the disentanglement on.

	dSprites	Shapes3D	Cars3D	SmallNORB	CelebA
Factors (# of values)	Shape (3) Scale (6) Orientation (40) Position X (32) Position Y (32)	Floor hue (10) Wall hue (10) Object hue (10) Scale (8) Orientation (15) Shape (4)	Elevation (4) Azimuth (24) Object id (183)	category (10) Elevation (9) Azimuth (18) Lighting (6)	40 attributes (2 for each)

FactorVAE Metrics Kim and Mnih (25) argues the BetaVAE score has the tendency to fail into a spurious disentanglement and proposes a new metric based on a majority vote classifier. Representations are obtained after the generated samples with only factor k fixed. Normalizing each dimension in representations in terms of standard deviation. Index of dimension with lowest variances of normalized representation and the factor index k is the input/output of the linear classifier. The accuracy of the classification is the FactorVAE score.

Mutual Information Gap Chen et al. (7) assumes the disentanglement model has the property that most information of one specific factor is contained in one dimension or a group of certain dimensions. The mutual information gap is the summation of the difference between the highest and second-highest normalized mutual information between a fixed factor and dimensions in representation. The formula can be illustrated as below:

$$\frac{1}{K}\sum_{k=1}^{K}\frac{1}{H_{z_k}}(I(v_{j_k}, z_k) - \max_{j \neq j_k} I(v_j, z_k)) \tag{5}$$

Where K is the overall number of ground truth factors. v is the latent representation and z_k is the factors of latent variables and $j_k = \arg\max_j I(v_j, z_k)$.

DCI disentanglement As Eastwood and Williams (11) suggests, the disentanglement is measured by the entropy of relative importance for each dimension in predicting factors. First, we have to know the importance of each dimension of the representation for predicting each factor. The importance is determined by a regressing model such as Lasso or Random Forest in the original DCI implementation (11) or Gradient Boosting Tree in DisLib implementation (36). We note the importance matrix R where R_{ij} is the importance of the i-th dimension in prediction the j-th factor. Then disentanglement score for the i-th dimension is defined as $D_i = (1 - H_K(P_i))$ where $H_K(P_i) = -\sum_{k=0}^{K-1} P_{ik} log_K P_{ik}$ denotes the entropy and $P_{ij} = R_{ij}/\sum_{k=0}^{K-1} R_{ik}$ denotes the normalized importance of i-th dimension in prediction the j-th factor. Finally the overall disentanglement score is calculated as $D = \sum_i \rho_i D_i$ where $\rho_i = \sum_j R_{ij}/\sum_{ij} R_{ij}$ is the weighting of the each dimension's informativeness in representing factors.

SAP (29) proposes the Separated Attribute Predictability (SAP) score. A score metrics is computed with classification score of predicting j^{th} factors on i^{th} dimension as the ij^{th} entry. SAP is the mean of the difference between the highest and second-highest scores for each column.

We follow the implementation provided by DisLib (36) for the evaluation protocol. Despite exceptions, the evaluation batch size is 64, the *prune_dims.threshold* is 0.06. If a classifier is required to be trained during evaluation, *num_train* is 10000, and *num_eval* is 5000. For Mutual information computation, the discretizer function is the histogram discretizer, and the number of bins in the discretization is 20. For the evaluation of MIG and SAP on dSprites, SmallNORB, Cars3D, and Shapes3D, BYOL representation vectors are reduced to 10 dimensions by PCA to be aligned with other methods. For the evaluation of MIG and SAP on CelebA, to have a fair comparison, the representation vectors of all methods are reduced to 40 dimensions. For the implementation of our proposed MED, the basic logic is the same as DCI Disentanglement, but we replace the classifier output with the mutual information based scores.

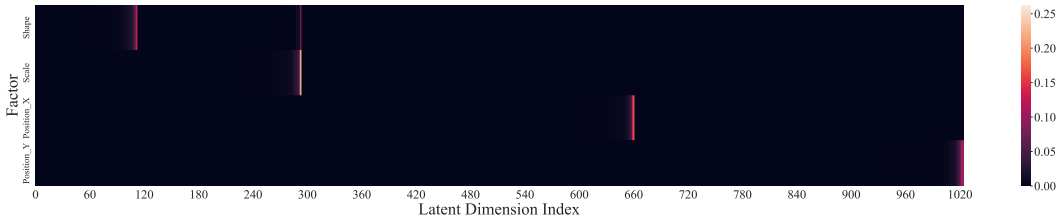
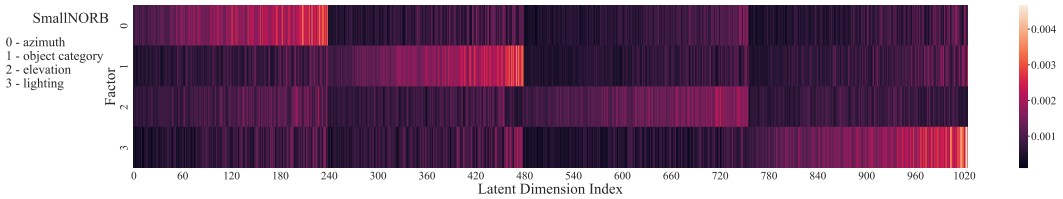
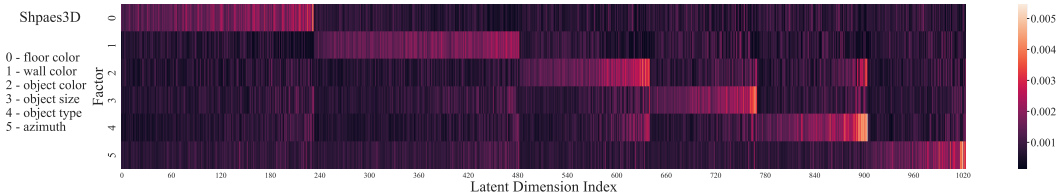


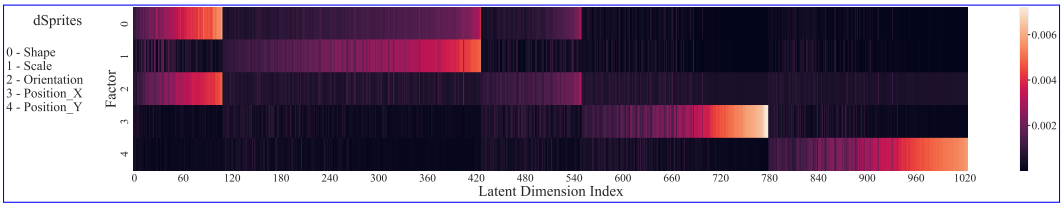
Figure 6: The importance distribution for the representation learned from BYOL on dSprites. Here, we follow the practice of DisLib to use a Gradient Boosting Tree (GBT) regressor to determine the importance matrix of each latent dimension in predicting each factor. Compared with the Mutual Information distribution shown in Figure 1a, the importance distribution is significantly more sparse. Sparsity is encouraged when constructing the GBT regressor. This makes it hard to study the true representation pattern.



(a)

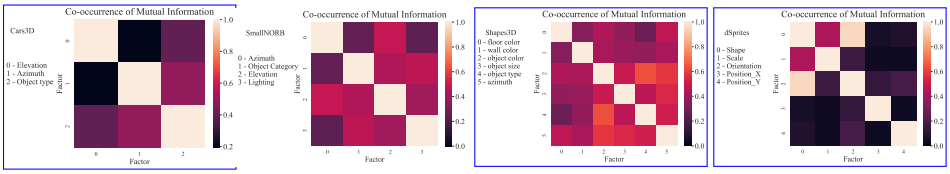


(b)



(c)

Figure 7: The mutual information distribution on SmallNORB(a), Shapes3D(b) , and dSprites including all factors(c).



(a) **(b)** **(c)** **(d)**

Figure 8: The co-occurrence of factors in the mutual information relationship among BYOL representations on Cars3D(a), SmallNORB(b) , Shapes3D(c), and dSprites including all factors(d).

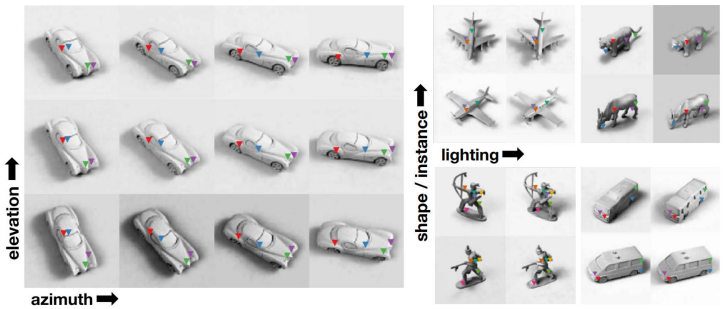


Figure 9: Some samples from SmallNORB dataset. The variance is controlled by the factor indicated on axis. The image is from Jakab et al. (21).

B More Qualitative Study

Limited by the main text length limitation, we provide more qualitative studies about the disentanglement property shown by the contrastive learning here. We still use BYOL as an example of the negative-free contrastive learning methods.

B.1 Importance Distribution by DCI

In the main text, we concisely talked about the potential variables introduced by the learnable model under some metrics. Here we show an example for the widely used DCI Disentanglement metric. We follow DisLib to use Gradient Boosting Tree to produce the Importance estimation between each factor and each latent dimension. All parameters are set the same as its default protocol. The visualization is shown in Figure 6. Compared with the mutual information distribution shown in Figure 1a, the importance distribution is obviously much more sparse. Sparsity is encouraged during constructing the GBT regressor. However, the observation can lead to the misunderstanding that the correlation between factors and latent dimensions is sparse which is not true. By using the pure measurement without involving additional adaptive models, such problem will not be raised in the proposed MED metric.

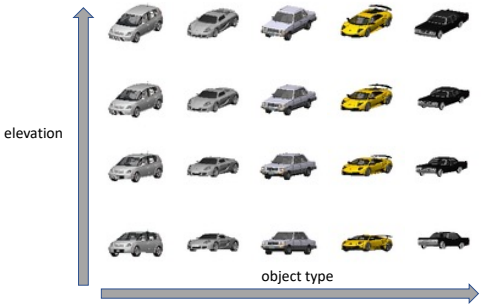


Figure 10: Samples from Cars3D. Object type and elevation are controlled. It show that the two factors are not independent.

B.2 Mutual Information Heatmaps

We compute MI between each latent dimension and each generative factor and visualize them by heatmaps, which offer us an intuitive picture of the learned representation space. For completeness, we show the MI heatmaps of SmallNORB, Shapes3D, and dSprites with all factors in Figure 7a, Figure 7b and Figure 7c respectively. We can see that the disentangled pattern described in the main text still emerges. There is a group of columns brighter than others in each row, and these groups do not overlap for most rows. However, we find that some latent dimensions may emphasize more than one factor. We provide a more detailed analysis from the perspective of factor co-occurrence on this phenomenon in sectionB.3 below.

B.3 Co-occurrence of Factors

To understand to what extent one dimension of the learned representation would respond to more than one factor, we make the co-occurrence of mutual information to factors on more datasets here. The visualizations are shown in Figure ,8b, Figure 8a and Figure 8c on SmallNORB, Cars3D, and Shapes3D respectively.

SmallNORB Though most non-diagonal entries have very low co-occurrence of mutual information, two pairs of factors show slightly higher co-occurrence. They are "azimuth-elevation" and "instance category-lighting". After investigating the dataset, we find the two pairs of factors are not fully independent. Figure 9 show some samples with corresponding factors manipulated. We could see that the elevation and azimuth are not fully independent. And the correlation between the instance category and the lighting factor is even more obvious because the lighting condition is sensibly related to the shadow around the object, whose distribution and shape are highly determined by the instance category.

Cars3D Only one pair of factors show some co-occurrence, i.e. "elevation-object type". We randomly selected samples from Cars3D by different object types and elevations, as shown in Figure 10. It shows that with the same value of elevation, samples of different object types have different visual elevations. So these two factors are not fully independent. This might explain the slightly higher co-occurrence of mutual information between this pair of factors.

Shapes3D The result shows relatively bad disentanglement. To be precise, some factor pairs show low mutual information co-occurrence as expected, such as the color factors of floor, wall, and object and the pair of "object color - azimuth". But the MI co-occurrence of "wall color - object size" and "object color - object size;" are higher than we expected as we did not recognize their high dependence. This result might relate to our model's relatively poor performance on Shapes3D.

dSprites We omit the factor *orientation* in the visualization at main text for clarity. Here we show the results including all factors. For dSprites dataset, the value of *orientation* is highly correlated with *shape*. This is because that different *shape* has different symmetry properties for rotation. As a result, dimensions that encode information of *shape* tend to capture information of *orientation* simultaneously. This explains that, in Figure 7c, some dimensions have two brighter entries at "shape" and "orientation" rows. Thus, it is not surprising that the "shape - orientation" cell is brighter than other off-diagonal elements in Figure 8d.

B.4 Manipulating Factors

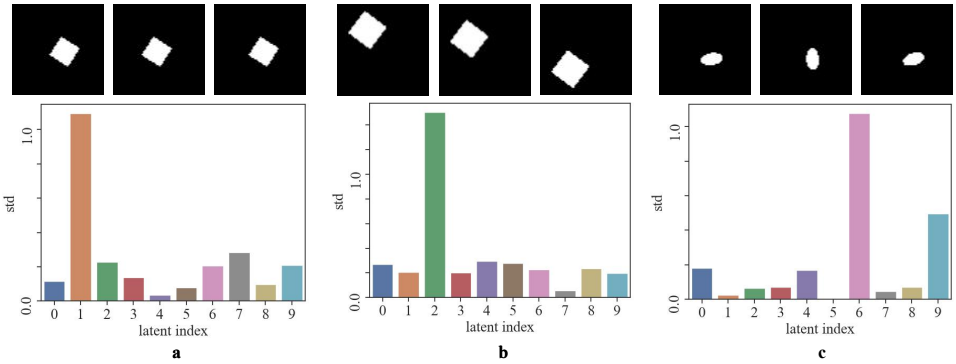


Figure 11: Representation variation when manipulating one factor only in the dimension-reduced version. In (a) and (b), *position_x* and *position_y* are manipulated respectively and only cause one dimension significantly variate. While, in (c), when manipulating the ill-defined factor *orientation*, two dimensions variate.

In the main paper, we studied the influence to representation by manipulating the factors, where the representation is reduced by selecting dimensions as in calculating Top-k MED. Here, we do the qualitative study of the influence on representation by manipulating factors in another way but still on dSprites. To make the original high-dimensional representation space more compact, we use the unsupervised dimension reduction by PCA instead, which is more general when the factor pattern is unknown. Here, we reduce the representation dimension by PCA to 10. Note that since the PCA operation mixes the original latent space with a linear combination, it might destroy the existing disentanglement property in the high dimensional space, or enhance the disentanglement if the original high dimensional space is a linear combination of the ground truth factors. But such influence is usually considered secondary to the disentanglement learned by a model. No matter

which case, if the dimension-reduced representation shows disentangled properties, the original space at least captures linearly transformed ground truth factors, and the dimension reduction techniques such as PCA can make the representation more compact in a qualitative study.

Figure 11 shows the result of representation vector variation when changing only one factor at once. Given three images with only one factor's value being different, we generate the 10-dim representation vectors from them. Then, we compute the variance across the three vectors, leading to 10 scalars. The larger the variance is, the more that dimension responds to the factor change. Figure 11(a) and (b) show how reduced representation vector changes when manipulating *position_x* and *position_y* factor respectively. It shows good disentanglement that only one representation dimension has high variation. However, in Figure 11(c) we show a failure mode of the ill-defined factor *orientation* that change of factor causes both the 6th and the 9th dimensions of reduced representation to have large variations. From the results, we observe that manipulating one well-defined independent factor causes evident variance in only one dimension. And it shows that we could make the learned representation vector more compact by unsupervised dimension reduction.

C More Quantitative Results

We present the complete disentanglement benchmark results in the next section because of length restrictions. Moreover, we plot the MED version of Figure 5 to reveal the significance of the latent dimension for disentangled representation learning.

C.1 Full Disentanglement Benchmark

In the main text, we evaluate the disentanglement under our proposed MED metric for full space and top-k subspace on multiple datasets. In this section, to provide a more complete understanding of the disentanglement property of contrastive learning without negatives, we report the disentanglement scores with other metrics, such as FactorVAE score, BetaVAE score, MIG, SAP, and DCI Disentanglement, here.

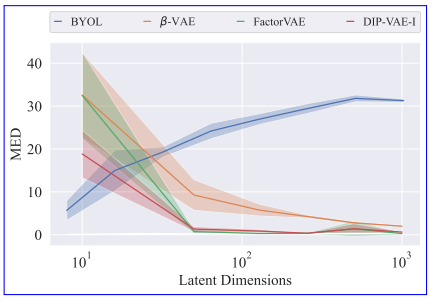


Figure 12: The influence of the representation dimension on the MED score on dSprites dataset. The overall disentanglement property is enhanced with increasing representation dimension for the BYOL method but decreased for all tested VAE-based methods. The minimum dimension is 10 for VAE-based methods, and 8 for BYOL

For the results of VAE-based methods, as the large-scale benchmark of Locatello et al. (36) provides the original logs on dSprites, Cars3D, and SmallNORB datasets, we simply report the performance of the best configuration. The original logs on Shapes3D are not available, so we train and evaluate on Shapes3D by ourselves for the MED scores. For scores under other metrics, we report the median disentanglement scores. Some results are from Locatello et al. (37) but the std error is not available. The median performance of SlowVAE is from its original paper (27). For the results of CelebA, the result of InfoGAN-CR is from its officially released checkpoint without availability to the std error. For other methods, we report the mean value of our trained weights over three random seeds as default. Because the evaluation of DCI is extremely time-consuming, around 14 hours for a 1000-d model, we only take BYOL as an example here for negative-free contrastive learning methods. All results are combined and shown in Table 4. We also include traditional metrics results on their selected top-k subspace for contrastive learning methods.

Same as the analysis we provide in the main text, the results show significant disagreement among the existing metrics. To be precise, for those metrics (BetaVAE score, FactorVAE score, DCI Disentanglement) using a learnable model such as a regressor or classifier, the high-dimensional BYOL model achieves a significant advantage. However, for the metrics relying on only one or two dimensions to reveal the connection between a latent dimension and a factor (MIG and SAP), BYOL's performance is not that impressive anymore.

Finally, the result on CelebA shows the great robustness of BYOL's learned representations to be disentangled on real-world datasets. Yet, the large gap between the score of CelebA and those on synthetic datasets emphasizes the difficulty of learning disentangled factors on real-world images. It

Table 4: Evaluation results on multiple datasets with different disentanglement metrics. The best results of all models are **bolded**. The best results on low-dimensional space are underlined.

	Model	BetaVAE	FactorVAE	MIG	SAP	DCI	MED
dSprites	β-VAE	82.3 (7.6)	65.8 (9.2)	26.3 (11.0)	5.2 (2.7)	39.3 (13.2)	32.6 (10.0)
	β-TCVAE	86.7 (2.4)	76.6 (7.8)	23.8 (6.8)	6.9 (0.9)	36.3 (7.1)	31.8 (7.4)
	FactorVAE	84.9 (2.8)	75.3 (7.4)	18.4 (9.0)	6.8 (0.8)	28.8 (10.6)	32.5 (10.1)
	DIP-VAE-I	82.7 (3.3)	59.1 (4.8)	9.6 (5.1)	5.2 (2.6)	14.4 (4.6)	18.8 (5.6)
	DIP-VAE-II	81.5 (4.9)	58.6 (7.6)	7.4 (3.4)	3.6 (2.2)	12.3 (5.2)	14.7 (5.5)
	AnnealedVAE	86.5 (0.1)	60.1 (0.0)	**35.2 (1.3)**	7.6 (0.5)	37.9 (2.1)	35.8 (0.8)
	Ada-GVAE	88.0 (2.7)	73.1 (3.9)	17.3 (4.7)	6.6 (2.0)	32.3 (4.6)	–
	SlowVAE	87.0 (5.1)	75.2 (11.1)	28.3 (11.5)	4.4 (2.0)	47.7 (8.5)	–
	EBM	82.3 (2.0)	65.7 (12.5)	1.7 (0.5)	3.0 (1.2)	19.1 (1.8)	6.8 (4.0)
	InfoGAN-CR	85.5 (1.0)	88.0 (1.0)	19.8 (3.2)	6.0 (1.0)	14.0 (5.2)	–
	BYOL	**93.2 (0.4)**	**91.6 (0.8)**	29.3 (0.4)	**8.0 (0.4)**	**66.9 (0.2)**	31.3 (0.4)
	BYOL (top-2)	90.0 (0.9)	79.2 (2.3)	3.3 (0.9)	0.8 (0.2)	45.0(0.1)	**53.7 (0.7)**
Cars3D	β-VAE	**100.0 (0.0)**	89.3 (1.2)	11.7 (1.1)	1.4 (0.9)	38.7 (4.6)	29.0 (2.2)
	β-TCVAE	**100.0 (0.0)**	92.2 (2.7)	**15.5 (2.9)**	1.7 (0.3)	42.7 (3.5)	**33.0 (3.8)**
	FactorVAE	**100.0 (0.0)**	91.7 (4.1)	10.6 (2.2)	**2.0 (0.5)**	29.0 (6.7)	29.1 (3.0)
	DIP-VAE-I	**100.0 (0.0)**	90.5 (5.0)	5.9 (2.8)	1.9 (1.4)	22.6 (5.6)	19.4 (3.3)
	DIP-VAE-II	**100.0 (0.0)**	85.0 (6.1)	5.1 (2.7)	1.3 (0.8)	20.8 (5.4)	16.7 (4.1)
	AnnealedVAE	**100.0 (0.0)**	85.0 (4.3)	7.6 (1.0)	1.5 (0.5)	18.5 (4.3)	15.5 (2.5)
	SlowVAE	**100.0 (0.0)**	90.4 (0.5)	15.4 (2.2)	1.6 (0.5)	48.0 (2.4)	–
	BYOL	**100.0 (0.0)**	**95.8 (1.2)**	7.6 (0.9)	1.8 (0.7)	**48.5 (2.3)**	9.7 (0.5)
	BYOL (top-3)	**100.0 (0.0)**	95.2 (0.8)	3.8 (0.5)	1.1 (0.8)	15.8 (3.6)	31.8 (1.3)
SmallNORB	β-VAE	84.1 (2.7)	60.1 (2.4)	25.0 (1.1)	11.4 (1.1)	32.6 (0.6)	24.4 (0.7)
	β-TCVAE	84.5 (2.7)	60.3 (2.3)	25.4 (0.9)	11.7 (1.1)	35.2 (0.7)	25.0 (0.9)
	FactorVAE	80.8 (3.8)	62.5 (3.6)	23.9 (2.0)	10.2 (0.9)	33.4 (1.1)	**25.9 (1.2)**
	DIP-VAE-I	84.2 (3.2)	69.8 (4.6)	24.3 (2.7)	10.2 (1.4)	30.0 (2.1)	24.5 (2.1)
	DIP-VAE-II	85.2 (1.3)	58.4 (2.1)	**25.5 (1.5)**	**14.4 (0.4)**	32.3 (0.7)	24.4 (0.7)
	AnnealedVAE	60.8 (6.2)	50.0 (9.9)	9.1 (2.2)	6.8 (0.8)	15.7 (6.4)	5.5 (3.7)
	SlowVAE	78.2 (3.8)	47.0 (2.9)	23.8 (1.8)	7.8 (1.1)	28.7 (0.7)	21.8 (1.3)
	EBM	79.0 (4.4)	57.9 (3.5)	1.7 (0.5)	1.9 (0.1)	13.9 (2.2)	2.3 (1.7)
	BYOL	**97.0 (0.8)**	**81.0 (0.5)**	3.3 (0.9)	2.2 (0.3)	**51.0 (1.0)**	7.7 (0.2)
	BYOL (top-2)	86.7 (0.4)	65.6 (3.7)	3.3 (1.4)	1.5 (0.2)	13.6 (0.3)	25.7 (0.3)
Shapes3D	β-VAE	98.6	83.9	22.0	6.2	58.8	12.9 (3.5)
	β-TCVAE	99.8	86.8	27.1	7.9	70.9	13.7 (0.9)
	FactorVAE	94.2	82.5	27.0	6.1	67.2	0.7 (0.9)
	DIP-VAE-I	95.6	79.7	15.2	4.0	55.9	10.3 (0.9)
	DIP-VAE-II	97.8	88.4	18.1	6.3	41.9	–
	AnneledVAE	86.1	80.9	35.9	6.2	47.4	–
	Ada-ML-VAE	**100.0**	**100.0**	50.9	12.7	94.0	–
	Ada-GVAE	**100.0**	**100.0**	56.2	**15.3**	**94.6**	–
	SlowVAE	**100.0 (0.1)**	97.3 (4.0)	**64.4 (8.4)**	5.8 (0.9)	82.6 (4.4)	–
	EBM	75.9 (11.2)	53.2 (8.7)	5.2 (2.2)	2.8 (1.1)	21.8 (11.0)	2.1 (2.6)
	BYOL	91.5 (3.9)	82.5 (2.4)	5.2 (1.7)	2.8 (0.3)	53.1 (1.5)	6.0 (0.5)
	BYOL (top-2)	95.5 (1.1)	83.0 (3.6)	2.8 (0.8)	1.4 (0.5)	27.2 (3.8)	**19.7 (1.3)**
CelebA	VAE	21.5 (3.2)	6.1 (3.8)	0.8 (0.1)	0.9 (0.2)	11.2 (2.3)	3.8 (0.2)
	β-VAE	19.1 (1.9)	5.8 (1.8)	0.1 (0.1)	0.6 (0.2)	8.7 (1.9)	3.3 (0.1)
	β-TCVAE	19.9 (2.3)	9.8 (2.4)	0.6 (0.2)	1.2 (0.3)	3.5 (1.1)	4.7 (0.1)
	FactorVAE	25.3 (3.0)	**12.0 (2.1)**	0.4 (0.1)	0.6 (0.2)	7.1 (0.7)	0.6 (0.6)
	DIP-VAE-I	21.0 (1.9)	9.3 (1.1)	0.2 (0.1)	0.9 (0.3)	13.8 (2.2)	3.6 (0.2)
	InfoGAN-CR	16.8	11.3	1.6	2.8	22.0	–
	BYOL	**35.7 (2.1)**	11.5 (1.1)	**2.6 (0.7)**	**8.2 (0.9)**	**41.0 (1.3)**	4.8 (0.4)
	BYOL (top-3)	25.8 (0.8)	9.7 (0.5)	0.7 (0.2)	0.2(0.02)	15.8(0.3)	**6.8 (0.7)**

is hard to empirically study whether it is the high dimension that gives BYOL advantages on some metrics because the nature of BYOL makes it hard to be trained with a small latent dimension to make a comparison.

In the last rows for each dataset, we show previous disentanglement metrics scores on the top-k subspaces of BYOL. We find that BYOL (top-k) has performed better than or on par with unsupervised methods with low latent dimensions. Note that Ada-GVAE, SlowVAE, and EBM are (weakly) supervised methods. The SOTA performance justifies our claim of the well-disentangled subspaces. MIG and SAP require a large gap between the two most vital representation dimension-factor responses, measured by mutual information and classification accuracy. But out of 1000 dimensions, we choose the k most relative ones. Hence the gap ought to be minimal. Consequently, BYOL (top-k) receives lower MIG and SAP scores. Except for dSprites and CelebA, BYOL (top-k) have lower DCI scores. The primary cause of this is the top-k process's massive information loss, which completely obliterates the original encoding pattern. As a result, the gradient boosting tree (GBT) developed throughout the DCI process tends to "borrow" information from latent dimensions that emphasize additional relative factors to classify some complex factors. The "information-lending" dimensions then become less disentangled since they are also important for predicting those complex factors. Taking the Cars3D dataset as an example, we set $k = 3$ for top-k MED. To predict the factor *object type*, the GBT needs to classify 183 types of cars from only 9 dimensions with only 3 emphasizing *object type*. This forces GBT to utilize information from *elevation* since these two attributes are correlated (see Appendix B.3). However, we keep all they have learned from training for other low-dimensional methods. Therefore, the information loss problem does not apply to them.

In summary, we obtain the following conclusions regarding metrics and top-k subspace with the complete benchmarks. **(1)** The scores of the whole space of BYOL empirically verify the disagreement of previous metrics. **(2)** Traditional metric scores on the top-k subspace support the well-disentangle property of the subspace.

C.2 Infuence of Dimension on MED

In the main text, we show the influence of dimension on top-k MED. To make the conclusion more rigorous, we report the correspondence analysis for MED in Figure 12. Combining with Figure 5, we cover disentanglement performance for BYOL with lower dimensions and VAEs with higher dimensions. We find that as the dimension grows, the MED scores evolve in the same pattern as top-k MED scores. This suggests that the dimension of representation influences both overall and partial disentangle properties. Thus representation dimension should be considered a core parameter for disentangled representation learning.

D Ablation Study

Limited by the main content page length, we put some additional ablation studies here to help better understand the influence of important inductive bias of BYOL when studying representation disentanglement.

Table 5: Results of using different normalization strategies on dSprites.

normalization	w/o norm	BN	GN	LN	IN
MED	23.8 (0.6)	29.4 (0.5)	31.3 (0.4)	31.3 (0.8)	0.0 (0.0)

D.1 Normalization

We experiment with five normalization layers configuration in the encoder network on the dSprites dataset. The results are shown in Table 5. For group normalization, we set the group number to 4. On dSprites, we find the commonly used BN decreases the disentanglement performance. By keeping the batch norm in the projector and the predictor, removing the batch norm in the encoder will not cause the model to collapse, which agrees with the observation in previous works (42). On the contrary, replacing batch norm in encoder with group norm or layer norm will increase the representation disentanglement while achieving similar accuracy in downstream factor prediction. We notice that a similar phenomenon has been discovered before in supervised representation disentanglement. For example, Bau et al. (3) discovered that a network trained with batch normalization layers has less interpretable (disentangled) neurons. On the other hand, instance norm (50) completely breaks the

contrastive learning process. We still do not fully understand this behavior, but we hypothesize that it may be caused by the shared batch statistics that make it hard for a feature to be aligned to the ground truth factor.

E Limitations

Our work still has some limitations, especially considering the design of contrastive learning methods still depends on heavy empirical practice. For the fair comparison in the benchmark, we use a shared encoder architecture for all methods but they may still have other inductive bias potentially influencing the results such as the hyperparameters in VAE-based methods and contrastive learning methods. We select the normalization in BYOL as an example in the ablation study above showing that such inductive bias can make influence over the evaluation results. But we can not do the ablation study on all possible inductive bias. We basically inherit the available best settings from DisLib (36) if possible and the settings from the public official implementations of other methods. All hyperparameters and details we customize have been indicated in the Section of Reproducibility.

F MED Details

The main text briefly introduces MED's motivation and evaluation process. In the following section, we will first explain why we can trust MED to measure disentanglement properties. Then we discuss the drawbacks of previous metrics in high dimensional space in detail. Finally, we provide more empirical results on the properties of the top-k subspace.

F.1 Intuition and Rationale

MED is designed to measure the *disentanglement* property following the DCI framework (11), i.e., the degree to which a representation dimension only responds to a single generative factor. The notion of disentanglement implies two requirements. The metric should **(1)** accurately measure the latent dimensions' "response" to factors, and **(2)** demonstrate the singleton correspondence to factors of latent dimensions.

To satisfy the above two properties, we combine two widely-accepted tools, mutual information (MI), and entropy. By definition, MI is a standard measurement of information shared by variables. Therefore the importance matrix calculated by MI (see Eq. 1) indeed indicates the correlation between representation dimensions and factors. As for the second requirement, we show that the *disentanglement* pattern matches the mechanism of entropy in 3.2 because it indicates the informativeness a dimension has for factors. And by normalizing entropy across factors, we measure the relative importance of one dimension to different factors, which fits the intention of estimating disentanglement.

Here we present some examples to verify. According to the row distribution defined as $P_{ij} = R_{ij}/\sum_{k=0}^{K-1} R_{ik}$, a perfectly disentangled dimension induces that $P_{ik} = 1, P_{ij\neq k} = 0$ for a certain k, yielding $S(i) = 1$ (see Eq. 2) . On the contrary, a fully entangled dimension captures combination information of all factors, corresponding to a uniform $P_{i'\cdot}$, yielding the minimum $S(i') = 1 - \log K$, where K is the number of ground truth factors. In summary, the two requirements naturally follow the definition of MI and entropy. Therefore MED is indeed a metric of *disentanglement* property.

F.2 Limitations of Previous Metrics in High-Dimensional Space

In 3.2 we report our observations of the flaws of existing metrics. Here we will discuss their limitations in detail following the two requirements described in F.1. In addition, we will show how MED can resolve the issues of previous metrics.

Metrics based on learnable classifiers, namely the **BetaVAE score** and **FactorVAE score**, fail to satisfy both of the requirements. As the main text describes, a higher dimension with redundant information makes it easier to overfit the classifier. As in evaluating disentanglement, standard protocol draws training and testing sample points from the same distribution. Such overfit can make a huge advantage in final scores. Therefore the classifier developed during evaluation tends to overestimate the "response". Thus their mechanisms are not aligned well with the *disentanglement* in high-dimensional cases.

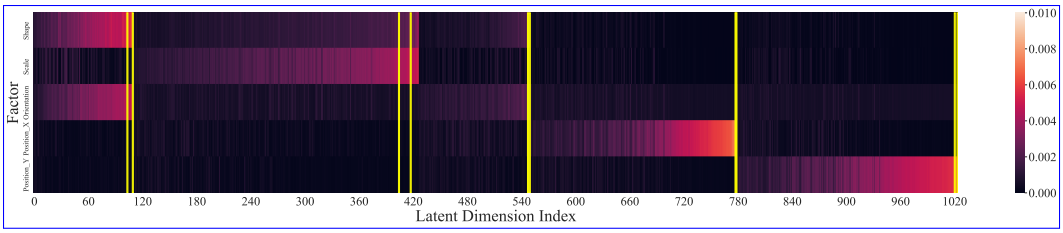


Figure 13: The visualization of the top-k process. Based on Figure 7c, the selected dimension are highlighted with yellow boxes. For dSprites we set $k = 2$. The indices picked are 103, 109, 404, 417, 547, 549, 777, 778, 1020, and 1023.

Table 6: *Logistic Regression* accuracy results of the full representation space and top-k subspace. Methods in gray are contrastive self-supervised learning methods.The best results of all models are **bolded**. The best results on low-dimensional space are underlined.

Space	Model	dSprites	Shapes3D	Cars3D	SmallNORB	CelebA
Full Space	β-VAE	34.5 (3.0)	35.6 (2.6)	65.0 (1.6)	48.6 (0.5)	82.1 (0.1)
	β-TCVAE	32.1 (3.4)	40.3 (2.1)	68.3 (1.4)	47.4 (0.7)	**85.2 (0.1)**
	FactorVAE	30.4 (2.5)	20.7 (2.3)	56.1 (1.3)	42.6 (1.3)	80.4 (0.2)
	DIP-VAE-I	30.3 (0.9)	37.6 (1.5)	68.5 (1.2)	49.6 (0.5)	**85.2 (0.1)**
	DIP-VAE-II	30.9 (1.9)	–	67.2 (1.7)	48.5 (0.5)	–
	MoCo	46.3 (1.3)	78.3 (2.1)	61.8 (1.4)	56.3 (0.7)	**85.2 (0.3)**
	MoCov2	52.0 (0.3)	76.5 (0.3)	63.1 (1.2)	53.8 (0.5)	84.9 (0.1)
	BarlowTwins	41.7 (0.2)	74.5 (0.8)	79.0 (0.2)	**62.0 (0.1)**	83.1 (0.2)
	SimSiam	40.3 (0.1)	**80.9 (3.4)**	45.3 (4.6)	54.5 (0.5)	84.0 (0.2)
	BYOL	**54.6 (0.2)**	71.5 (1.3)	**81.2 (0.4)**	57.2 (1.1)	81.9 (0.2)
Top-k Subspace	MoCo	22.7 (1.4)	43.7 (1.1)	38.1 (1.2)	34.5 (0.2)	85.1 (0.2)
	MoCov2	26.0 (1.0)	33.4 (0.2)	32.6 (0.8)	31.4 (1.1)	84.9 (0.1)
	BarlowTwins	29.1 (0.7)	44.4 (1.3)	43.1 (1.4)	34.5 (0.4)	82.9 (0.2)
	SimSiam	38.0 (0.6)	61.8 (1.0)	29.4 (3.6)	37.7 (2.5)	83.8 (0.2)
	BYOL	37.9 (0.6)	46.3 (0.5)	43.3 (0.9)	40.2 (0.8)	81.9 (0.2)

Metrics like **MIG** and **SAP** are based on an extended intention of explaining *disentanglement*: not only should a dimension correlate to only a single factor but a factor should be responded by only a single dimension. In fact, they indicate the degree to which a factor is only captured by a single representation dimension, which is the *completeness* property in the DCI framework (11). For contrastive learning methods, the dimension of representations (about 1000) is much larger than the number of factors (less than 10). This increases the conflict of *completeness* and *disentanglement*. In more general cases where the factor number is unknown, these metrics would give a bonus to the models that happen to have dimensions close to the underlying factor numbers.

For **DCI Disentanglement Score**, the first prerequisite is also not met. The learnable regressor such as GBT (see Appendix A.2) tends to underestimate information captured by some dimensions since it encourages sparsity. The learnable module in DCI also makes unfair comparisons when the model representation dimension varies.

MED avoids the problems above by the following properties. Firstly, MED calculates MI between factors and representation dimensions one by one. Thus it is robust to dimension growth and MI is fair for different model dimensions. Moreover, MED focuses on the entropy of the rows of R, decoupling *disentanglement* and *completeness*.

F.3 Properties of the Top-k Subspace

We propose the top-k selection process to study the partial disentangle properties. Here we provide more details. We first visualize the top-k selection process based on the heatmap of importance

matrix. Then we will discuss the effects of k and the downstream performance comparison of the top-k subspace and full space.

Visualization of the top-k Process We describe the top-k process in the Section 3.3. Here we colored the picked dimensions yellow in the Figure 13. It shows that most picked dimensions only have one entry much brighter than the others, indicating their disentangled properties. Since *orientation* is ill-defined in dSprites, BYOL tends to encode it with *shape*. Hence the most disentangled dimensions for *shape* also capture information of *orientation*.

Ablation on k The parameter k is designed as an evaluation choice. For analytical purposes, we choose the value of k to match the dimension of VAE methods in the main text. To investigate the influence of k, we set $k = 1, 2, 10, 50, 100$ and evaluate the top-k MED scores of 1000-d VAEs and BYOL on dSprites. The results are shown in Figure 14. We find that with different values of k, the ranks of top-k MED scores remain unchanged. Since the top-k process is a supervised greedy selection, as k grows the scores decrease. In conclusion, top-k MED can measure the subspace disentangle properties with a wide range of k. Therefore we can choose k according to actual needs.

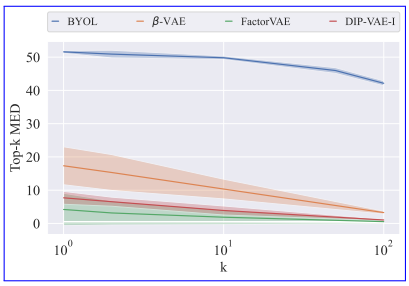


Figure 14: The influence of k in top-k MED on dSprites dataset. The evaluation results are consistent with multiple choices of k.

Downstream performance of top-k subspace We study the disentangle properties of the full space and top-k subspace in 5.3. Here we analyze the downstream performance of these spaces. We train a *Logistic Regressor* to predict each factor from the full representation or the top-k picked representation. Then we take the mean prediction accuracy of all factors as the measure of downstream performance. We compare the downstream performance of full space of VAEs and BYOL, and top-k subspace of BYOL on multiple datasets. Here we set $k = 2$ for dSprites, Shape3D, and SmallNORB, and $k = 3$ for Cars3D and CelebA to match the dimension of VAEs. The results are in Table 6. We note that, except on CelebA, performance drops non-negligibly when completing downstream tasks with the top-k representations. This suggests that the top-k subspace is well-disentangled but other dimensions also contain factor-related information. It is unsurprising because we cut down the dimension drastically (from 1000 to 10 on synthetic datasets and from 1000 to 100 on CelebA) by subscribing. However, we find that the top-k spaces learned by contrastive methods achieve better or compatible performance compared with the full spaces of VAEs on 3 out of 5 datasets (dSprites, Shapes3D, and CelebA). It is reasonable that a high-dimensional model can be better for downstream tasks as the downstream model can be trained to leverage each little piece of information from the input representations. But we get a message from the results that by drawing a subspace from the powerful CL-trained model, its performance on downstream tasks can be comparable to or superior to the low-dimensional models. And the selected low-dimensional model can have much better compactness and interpretability.

G Disagreement of existing Metrics

Limited by the page limit, in the main paper content, we can only briefly explain the disagreement of existing disentanglement metrics in Section 5.2. Figure 4 shows the disagreement of FactorVAE score, BetaVAE score, MIG, SAP, and DCI Disentanglement when the high-dimensional representation model joins the comparison. In the background that there is no uniformly recognized definition of "disentanglement", the agreement of existing metrics and their "basically shared" definition or expectation of "disentanglement" is one of the foundations of quantitatively evaluating disentanglement in this area since DisLib (36). However, the results in Figure 4 break the belief. Given the fact that self-supervised representation learning always requires a high dimension to be trained, we think the proposal of MED is necessary if we want to extend the study of disentangled representation learning to complicated real-world datasets and self-supervised representation learning methods.

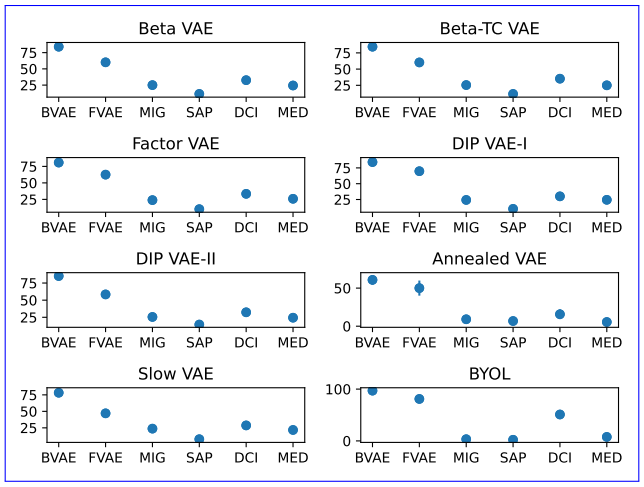


Figure 15: The re-organization of Figure 4. For some data points, the standard error of result is too small to present a long enough error bar stick in the figures. We note this figure may make confusion as the values of different metrics are usually at different scales and can't be compared directly. To more clearly understand the disagreement of existing metrics, we suggest referring to the raw results in Table 4 or Figure 16 below.

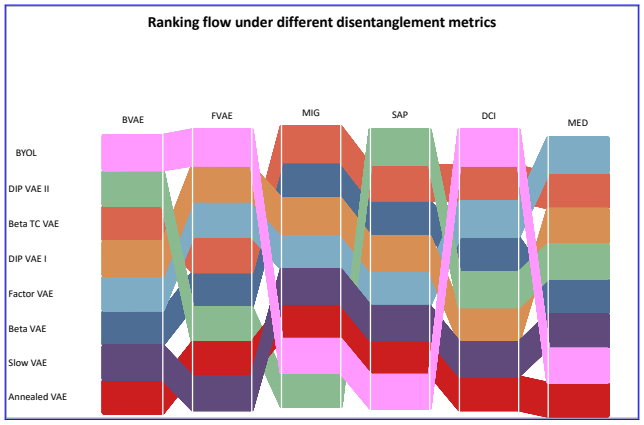


Figure 16: The flow chart of rankings of disentanglements scores. From top to down, the ranking decreases, indicating high to low disentanglement properties. If all metrics agree perfectly, there should be NO switch of relative ranking among these methods with different metrics. However, we notice obvious ranking variance, especially regarding the ranking of BYOL. With the analyses of the bias of existing metrics when the representation dimension varies, this figure emphasizes the necessity of proposing a new disentanglement metric that is fair to compare models of different dimensions.

Here, we redraw that figure with the value of multiple metrics on the same model in a subfigure in Figure 15. Moreover, we note by including multiple metrics of different meanings of scale into one subfigure, Figure 15 may cause confusion. So we make a visualization to show the disagreement of existing metrics by ranking different methods with them. The results are shown in Figure 16. If all metrics agree perfectly, there should be no relative ranking switch. However, the frequent switch, especially with BYOL being taken into comparison, strongly suggests their disagreement. Results in Table 4 further show that different metrics may make completely different rankings with large value gaps. All these results in this section show the obvious failure of using existing metrics to evaluate high-dimensional representation models. Together with the logistic analysis in Appendix F and the quantitative results in Appendix C, these results suggest the necessity to propose a new metric like MED to make it possible to introduce CL into the disentanglement benchmarks.